

GraDE: A Severity-Graded Dashcam Event Dataset with Per-Video Adaptive Segmentation

Mittal Darji^{1*}, Nikhil Gondaliya²

¹Research Scholar, CVM University, Vallabh Vidyanagar, Gujarat, India.

²Professor and Head, Department of Information Technology, G H Patel College of Engineering & Technology, CVM University, Vallabh Vidyanagar, Gujarat, India.

E-mail: ^{1*} mittaldarji54@gmail.com, ²nikhil.gondaliya@cvmu.edu.in

Orcid ID: ^{1*} 0000-0001-7840-5831, ²0000-0001-7011-5469

Abstract

The ego-view traffic datasets generally formulate crashes as a binary recognition problem, in which near-misses are largely excluded and collisions are not graded based on their physical severity. GraDE (Graded Dashcam Events) aims to fill this gap by providing a total of 1,321 videos collected from 25 public dashcam compilation uploads. Videos are categorized into three classes: Non-Crash, Near-Miss, and Crash. Among the 487 Crash videos, 433 have a Minor or Major grade indicating their severity as seen in the video footage. A second annotator rated a predefined set of 297 videos; Cohen's kappa was 0.656 for the 243 videos that neither of the two annotators considered ambiguous. In the compilation videos, unrelated events are spliced together by an editor. To select the clips from compilation videos, we utilized QualityGridAuto, which searches fourteen candidate distance thresholds for each source video and evaluates their quality based on usable duration, boundary strength, within-segment similarity. It applies adaptive thresholding to the construction of the dataset rather than developing a new segmentation method. The frames were sampled at 1 fps and embedded at 224×224 pixels, while each baseline utilized sixteen frames with the same resolution. In the manual audit, 93.2% of inspected clips contained one continuous scene. With source-pure four-fold testing, the CNN + BiLSTM baseline reached macro-F1 0.8725 on the three event classes and Video Swin-T 0.9053, but CNN + BiLSTM fell to 0.6889 when the two severity grades were added. On dedicated Minor versus Major classification the two models reached 0.5987 and 0.6510, so severity grading remains substantially harder than event classification. GraDE is released as an annotation index.

Keywords: GraDE Dataset, Dataset Construction, Dashcam Video, Near-Miss Detection, Crash Severity Grading.

1. Introduction

The safety triangle approach views major crashes as the tip of a much larger triangle whose base consists of dangerous encounters, most of which end without any collision occurring and therefore are never formally documented [1]. Naturalistic studies proved the base data to be useful. The instrumented fleets in the 100-Car study recorded near-crash events at higher rates

* Corresponding Author

Journal of Trends in Computer Science and Smart Technology, December 2026, Volume 8, Issue 4, Pages 1018-1045

DOI: <https://doi.org/10.36548/jtcsst.2026.4.006>

Received: 17.08.2026, received in revised form: 21.09.2026, accepted: 13.10.2026, published: 27.10.2026

© 2026 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

than actual crashes. The near-miss situations provided valuable insight into the causes of car crashes [2]. However, such research can only be done with small sample sizes due to equipment costs, limiting studies to several hundred cars in a single location. With the advent of consumer dashcams, a new source of video footage was added. Drivers make continuous recordings, most of which are subsequently uploaded to websites. Some channels compile these video clips into long videos with showcasing multiple events. Thus, an unstructured yet varied resource is formed. Such videos have already been used to generate accident datasets [3, 4]. The issue is that the dataset is not organized. A compilation video contains multiple events in one file, thus, it has to be segmented into clips, and then labeled.

Three limitations have led to this research. First is labeling. Ego view data usually define the problem as accident or not, sometimes mentioning the time of impact [3, 4, 5]. Near miss events are often ignored or labeled negative, although naturalistic evidence shows that near misses are frequent and might be useful for safety analysis [6]. The second limitation relates to the lack of a severity metric. Studies of traffic safety use police reports on injuries, which are defined after an event occurred. Video datasets should include an attribute that can be derived visually and is different from the severity of injury. The last limitation is the way clips are formed. Earlier techniques usually include sampling frames, computing distances between their representations, and using the same threshold value throughout the whole dataset [7]. In reality, the distance pattern varies depending on the speed of editing, overlay effects, branding, resizing, letterboxing, and camera motion. Consequently, a single threshold may be effective for one video but detrimental for another, leading to event splitting in the former case and merging in the latter. This study is based on the use of the GraDE dataset and the approach we have applied to develop this dataset. In Section 1, we introduce this research in terms of ego-centric video event datasets, edited video segmentation approaches and traffic severity assessment, demonstrating the drawback of each of them. In Section 2, we discuss the data collection process, the way we carried out video segmentation, refinement and evaluation of the resulting clips, assessment of their severity and the process of data separation. Section 3 provides more information regarding the dataset itself, threshold setting, the segmentation validation, events and severity benchmarks within the source-grouped release, and the limitations of the resource.

The specific contributions are as follows:

- GraDE provides 1,321 ego-view clips reconstructed from public dashcam compilations and labelled as Non-Crash, Near-Miss, or Crash. Human reviewers additionally assigned an apparent-physical-severity grade of Minor or Major to 433 of the 487 Crash clips.
- The release includes a severity rubric based on physical evidence visible in the video rather than injury outcome. It specifies how borderline cases are resolved and how clips with an obscured or incomplete impact are handled.
- QualityGridAuto replaces a corpus-wide segmentation threshold with a discrete per-video search scored from statistics of the resulting segments. It utilizes existing pipeline embeddings and requires no additional training. We consider this a dataset-construction application of established adaptive thresholding, not a new segmentation algorithm (Section 1.2). Its specific addition is a label-free rule that selects each upload's threshold based on the clips it would produce; its benefit lies in improved clips and boundaries, not better classifiers.

- The released four-fold protocol groups clips by source upload, preventing an upload from appearing in both training and test sets within the same fold. The release also contains the information required to reconstruct each clip.
- Reference benchmarks measure how well the three event classes can be learned and how much harder the graded-severity task is. Per-fold and per-class results, together with pooled confusion matrices, document where the dataset is easy and where errors remain.

1.1 Ego-View Traffic Event Datasets

The first widely adopted ego-view accident dataset is the Dashcam Accident Dataset (DAD), which defined the accident versus normal binary problem formulation [4]. DoTA, introduced by Yao et al., includes anomaly windows and categorical reasons for first person driving videos [8]. The Car Crash Dataset (CCD) [3] expanded the positive class with new attributes like lighting conditions, weather conditions, and ego-vehicle involvement. DADA-2000 combined accident videos with driver eye gaze recordings [9]. Later datasets have increased in terms of scale and attributes. MM-AU [10] is the largest dataset, including object-level annotations and reason-prevention text pair annotations. The Nexar dataset [5], uses fleet videos annotated for anticipation, and both collision and near collision videos are considered a single positive class rather than being benchmarked separately. This is due to the subjective nature of the alert moment, which required multiple annotators. Surveys provide the current state-of-the-art: Rocky et al. [11] classify dashcam techniques according to the type of supervision scheme employed. However, what has not been provided is an example of a severity score target defined on the video itself, a gap that GraDE aims to fill. In the meantime, the field has moved beyond the use of frame-level CNNs to utilize spatio-temporal video transformers [12], VideoMAE-like self-supervised video foundation models [13], and multimodal large language models for traffic videos [14, 15]. While GraDE employs a frozen ResNet-50 network as a distance descriptor, the purpose of the classifier is to provide a reproducible baseline for the dataset, not a survey of modern video architectures.

Table 1. Capability comparison of ego-view traffic event datasets. Y present, N absent; Combined = collision and near-collision annotated as one positive class. Index = identifiers released without video; Severity = graded collision outcome label, not an anomaly category

Dataset	Near-miss class	Severity grade	Scope of severity	Clips released	Source-disjoint protocol
DAD [4]	N	N	-	Yes	N
DoTA [8]	N	N	-	Yes	N
CCD [3]	N	N	-	Yes	N
DADA-2000 [9]	N	N	-	Yes	N
MM-AU [10]	N	N	-	Yes	N
Nexar [5]	Combined	N	-	Yes	N
Jaradat et al. [7]	Y	N	-	No	N
CycleCrash [16]	Y	Y	Cyclist only	Index	N
GraDE (ours)	Y	Y	General vehicle	Index	Y

Survey papers provide information about state-of-the-art. For example, Rocky et al. [11] classify methods for dashcams according to their supervision type. However, the current literature lacks an example of a target severity score defined directly on video. GraDE tries to fill this gap. In parallel, the current area has moved forward from CNNs at the frame level

to spatiotemporal video transformers [12], self-supervised VideoMAE-like video foundation models [13], and multimodal large language models for traffic videos [14, 15]. Despite the use of the frozen ResNet-50 network as a distance descriptor, the goal of the classifier is to develop a reproducible baseline for the dataset. The dataset most closely related to the present one in construction is that of Jaradat et al. [7], which also derives three-class clips from public dashcam compilations using cosine-distance segmentation of ResNet embeddings. It differs from the present work in three respects: the segmentation threshold is fixed rather than selected per video, no severity grading is applied within the crash class, and the resulting clips are not released for public use. The present work addresses each of these points.

Drawbacks Common to Prior Corpora: Two properties recur. The dominant label structure is binary; where a third class exists it denotes an anomaly type rather than a graded outcome. The same holds for the architectures built on these corpora: each targets whether an accident occurred, not how severe it was. A recent lightweight detector couples detection with a coarse severity label [17], but grades from a fixed-camera viewpoint and releases no ego-view clip set. Severity, when it appears, is confined to a specific road user (as in cyclist-collision datasets), or to synthetic and roadside rather than ego-view footage. In addition, DAD, CCD and the closest prior work of Jaradat et al. [7] derive their clips from compilation video without releasing the extracted clip set, so a downstream user cannot use the same clip pool. Table 1 summarises the comparison. All listed datasets are ego-view; DAD, CCD and the present work derive from compilation video.

1.2 Segmentation of Compilation Video

Partitioning edited video into shots is a mature problem in multimedia retrieval with a long evaluation history [18]. Classical detectors threshold colour-histogram or edge dissimilarity between adjacent frames, and learned detectors such as TransNet V2 predict transition probability from short frame windows [19]. Both aim to detect the shot instead of the event. New review articles have shown that the field is still active and that adaptive thresholds and not global thresholds are currently state of the art: Thounaojam et al. [18] provide a review of the problems and solutions in shot boundary detection and show that feature integration enhances the detection but increases computational costs. This literature proves that adaptive thresholding has become a recognized method, and this is exactly the reason why this project asserts the novelty of application for Section 2.4.

Limitation: A shot boundary is not necessarily an event boundary. A compilation may replay one incident or show it from several angles, while a continuous driving scene may have no editorial cut. For this reason, general shot detectors do not match the units needed by event datasets, and earlier dataset pipelines have used embedding and threshold rules [3, 7]. Those studies do not look at whether one threshold should cover the corpus or be chosen for each source video. We adopt the latter approach in Section 2.4. Section 3.4 tests it against manually marked incident boundaries in eight source videos and compares it with TransNet V2. Extending the manual boundary set to the complete corpus is left for future work.

1.3 Severity in Traffic Safety Analysis

In traffic-safety research, severity usually refers to injury outcome recorded by police or hospitals on an ordinal scale. Such analyses depend on retrospective records and record linkage, rather than on visual sensing. Video research has therefore focused mainly on detection

and anticipation [11]. When severity is predicted, the inputs are generally structured crash records: Santos et al. [20] review machine-learning methods for injury-severity prediction, while Chen et al. [21] study imbalance in injury-coded data. Near-crash studies have instead used instrumented vehicles and kinematic triggers [2], or extracted incidents from trajectories recorded by event systems. Large-scale near-miss analysis at fixed sites likewise relies on trajectories and surrogate indicators, not severity-annotated ego-view clips. Nevertheless, the perception systems used in this area require clip sets and therefore inherit the segmentation problem considered in Section 2.

Position Adopted in this Work: A dashcam does not reveal injury outcome, which also depends on restraint use, occupant characteristics, vehicle protection, and medical response. GraDE therefore records apparent physical severity: the magnitude of the event that can actually be seen in the video. It is not presented as a measure of injury severity. The operational criteria are given in Section 2.8.

2. Method

The pipeline converts a list of public video URLs into a partitioned, labelled, clip-level dataset through eight stages: source acquisition, uniform frame sampling and embedding, boundary hypothesis generation, per-video threshold selection, boundary refinement and cutting, manual review and labelling, severity grading of the Crash class, and source-aware partitioning. Figure 1 shows the overall flow.

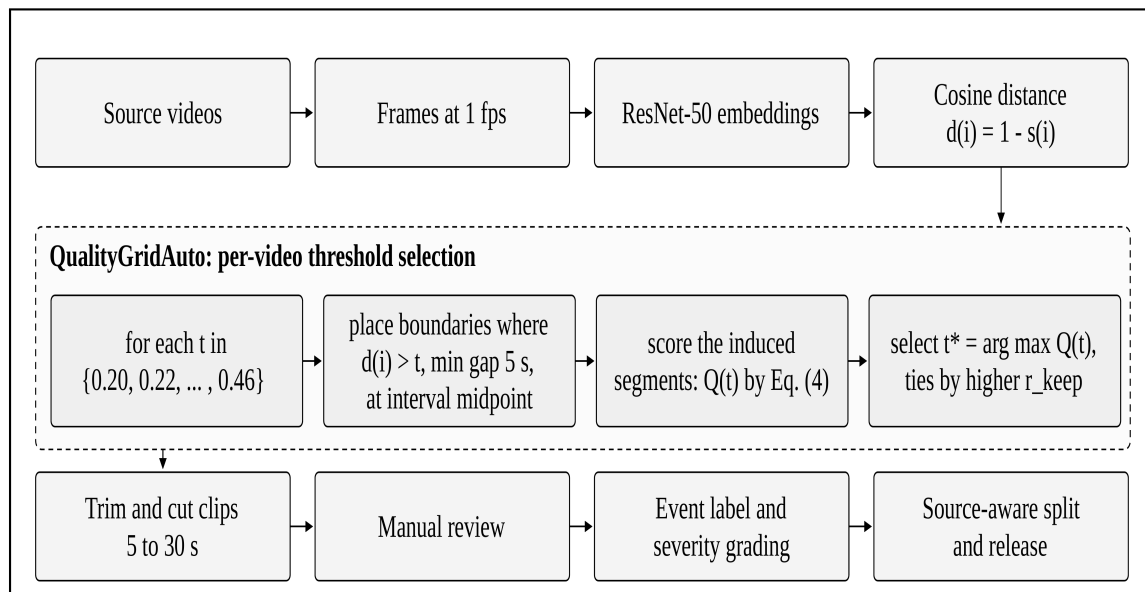


Figure 1. Clip generation and dataset preparation pipeline. QualityGridAuto (dashed box) selects the distance threshold independently per source video

2.1 Source Acquisition

We located source uploads by searching public platforms for channels publishing event-specific compilation videos. Search phrases included ‘dashcam crash compilation’, ‘near miss compilation’ and ‘dashcam driving footage’. Eligible channels had uploads that were thematically consistent and available at 480p or higher. Each selected video was added to a control file with a sequence number, URL and a provisional category. That category provided the label for the clips that came from it, but reviewers updated the label if the clip content did

not match the source category.

The pipeline takes in a chosen set of sequence numbers which allows new sources to be processed in batches. Each video is downloaded in the highest available progressive MP4 quality. It is accepted only if it passes checks for frame rate, number of frames, size and if it can be opened without issues. A stable source identifier is made by combining the sequence number with the platform's video ID. This same identifier goes with every clip that comes from that source, so the origin of each clip is always known.

2.2 Frame Sampling and Embedding

Each source was then decoded directly into JPEG frames at 1 fps; for these sources, it was more accurate to filter on frame rates rather than seek within containers. The first six sources had already been processed using OpenCV to take every round(fps)-th frame decoded, which also provides one frame per second. The quantity sampled is the transition within the editing between incidents, not the sub-second transitions that occur during a collision itself. Since the cut within editing defines the boundary between two scenes lasting many seconds each, 1 fps ensures that the number of descriptors will be proportionate to the length of video while still capturing the appropriate transition. The frame rate will not capture extremely short transitions within scenes, which lie beyond the scope of the intended boundary definition. This is also true of highway collisions at high speed: increasing the frame rate will provide better timing information within an incident but not significantly shift the cut between incidents. We tested this by re-running the complete pipeline at 1, 2 and 4 fps on the eight annotated videos, changing only the sampling rate. On the six videos sampled with the ffmpeg filter, where the 1 fps re-run reproduces the released boundaries exactly, boundary F1 within a 1 s tolerance was 0.941 at 1 fps, 0.839 at 2 fps and 0.838 at 4 fps (the marks were placed on 1 fps candidates, so a 1 s tolerance avoids penalising finer placement). Of the 103 marked cuts, 4 fps found 2 that 1 fps missed and missed 14 that 1 fps found, at four times the embedding cost. Higher rates spread gradual transitions over several frame pairs (marked cuts whose strongest pair fell below a distance of 0.30 rose from 9 to 16 of 126 at 4 fps) and added short within-scene spikes, so with the grid of Section 2.4 a finer rate did not recover missed boundaries; the grid was not re-tuned for other rates. On the two OpenCV-sampled videos, ffmpeg sampling changed the selected threshold (0.38 to 0.22 and 0.36 to 0.30), so on short uploads the selection can depend on the decoder.

Every sampled frame was resized to 224 by 224 pixels, converted to RGB colour and normalised with ImageNet channel statistics [22]. A ResNet-50 model that was loaded with ImageNet weights then produced a descriptor after its classification head was swapped out for a simple identity mapping [23]. The ResNet-50 network stayed fixed during this phase; it was not fine-tuned. No gradients were computed. In this context, ResNet-50 acts as a broad visual descriptor for measuring distances in the embedding space just like earlier compilation-based pipelines [3, 7]. The whole process relies on pairwise distances between descriptors, so any other general-purpose encoder could step in instead. Descriptors were saved separately for each source together with frame identifiers, timestamps and row positions so future runs can reuse them.

2.3 Boundary Hypothesis Generation

Let a source video yield descriptors f_1 through f_n at one-second intervals. For each adjacent pair the cosine similarity is computed on L2-normalised descriptors, and the dissimilarity used throughout is its complement, as given in Eq. (1) and Eq. (2).

$$s(i) = \frac{f_i \cdot f_{i+1}}{\|f_i\| \|f_{i+1}\|} \quad (1)$$

$$d(i) = 1 - s(i) \quad (2)$$

Large $d(i)$ signifies a significant variation in the visual elements, and the majority of such variations come from editorial cuts between incidents. An appropriate candidate for the boundary is considered to be index i if the dissimilarity value $d(i)$ surpasses a threshold value t and occurs no earlier than five seconds from the previous boundary. Such suppression is necessary since one dissimilarity value will signify several consecutive samples for a transition (such as a dissolve or a wipe) and will result in a number of nearly simultaneous boundaries. The five-second value is used as the target minimum value for the useful clip range (Section 2.4) and coincides with the minimum length of the event a downstream evaluator can make use of.

The boundaries occur at the middle of the span of frames that trigger them, which allows centering the cut in the temporal uncertainty window of one second caused by the sampling rate. The beginning of the video is always a boundary point; the end of the video is added if the last boundary occurred more than one second before it.

2.4 Per-Video Threshold Selection

The value of t controls the entire segmentation process. Fixing it once for a whole corpus presupposes that the distribution of distances is the same across source videos, which does not apply to crowdsourced corpora. This is because uploads vary in terms of editing speed, overlays, branding, scaling and camera ego-motion, each affecting the distribution. Hence, QualityGridAuto decides on the value of t separately for each source video. A set of 14 distance thresholds ranging from 0.20 to 0.46 with increments of 0.02 is considered. The lower boundary is chosen based on the knowledge that dissimilarities lower than 0.20 mean regular scene appearance variation in normal drive videos, while the higher one is chosen because dissimilarities higher than 0.46 are usually real editorial cuts. The step of 0.02 represents a compromise between coverage and evaluation cost. The procedure described in Section 2.3 is applied for each of these thresholds to create an assumed segment set. This set is then evaluated without cutting any clip, making the evaluation process less expensive, as the embeddings and distances between them are computed only once for all fourteen evaluations.

Scoring rests on three recognisable properties of a good segmentation: most segments should have a duration suitable for a training example, boundaries should fall at points of strong visual change, and segments should be internally coherent. Call a segment usable when its post-trim duration falls between five and thirty seconds. Define the usable, over-short and over-long proportions as in Eq. (3).

$$r_{keep} = \frac{|usable|}{|segments|}, \quad r_{short} = \frac{|short|}{|segments|}, \quad r_{long} = \frac{|long|}{|segments|} \quad (3)$$

Let $B(t)$ denote the mean dissimilarity value at the accepted boundaries, a measure of how decisively the segmentation cuts, and let $C(t)$ denote the mean cosine similarity among consecutive frame pairs lying strictly inside segments, a measure of how internally coherent the segments are. The composite objective is given in Eq. (4).

$$Q(t) = 0.40r_{keep} + 0.20B(t) + 0.20C(t) - 0.10r_{short} - 0.10r_{long} \quad (4)$$

The selected threshold is the maximiser of this objective, as in Eq. (5), with ties resolved in favour of the higher usable proportion and then in favour of fewer over-long and fewer over-short segments.

$$t^* = \operatorname{argmax} Q(t), \quad t \in \{0.20, 0.22, \dots, 0.46\} \quad (5)$$

The weights express a predetermined priority order. Usable-segment proportion receives the largest weight because clean cuts have little value if most resulting clips fall outside the allowed duration. Boundary strength and within-segment coherence have equal, smaller weights and capture different failure modes. A threshold that is too low creates repeated cuts at weak transitions, whereas a threshold that is too high combines visually unrelated material into long, incoherent segments. The short- and long-segment penalties are smaller because discarded segments primarily reduce yield rather than label quality. We therefore fixed the weights at 0.40, 0.20, 0.20, -0.10, and -0.10 before comparison with manual boundaries. Usable yield is weighted twice as strongly as either quality term, with equal small penalties on both duration failures. Section 3.3 tests this choice against random weightings and against weights tuned on the manual boundaries of Section 3.4.

The objective is only an unsupervised selection rule. It was not fitted to human boundary annotations and does not claim to estimate boundary accuracy. For each video it ranks fourteen candidate thresholds. Because usable yield has the largest contribution, the gain reported in Section 3.3 should be read as an improvement in the pool of candidate clips. Accuracy relative to human boundary marks is assessed separately in Section 3.4.

2.5 Boundary Refinement and Clip Cutting

We trimmed 0.50 s from the beginning and 1.25 s from the end of every segment. Material close to a cut can contain the previous incident, a transition, or a title card, and fixed symmetric trimming did not reflect how these artefacts appeared in the source compilations. The longer trailing removal was chosen because frames after the detected cut more often belonged to the next incident, channel bumper, or title card, whereas the preceding side usually contained only a short fade-out. The audit in Section 3.4 supports this direction: the 1.25 s trailing trim removed all following-scene material, while five clips retained content at the leading edge after the 0.50 s trim. We also measured edge contamination against the manual boundaries of Section 3.4, counting a clip as contaminated when a true cut lies inside it within 2 s of either end. With 0.50/1.25 s none of the 105 usable clips in the annotated span was contaminated; without trimming, 11.8 and 5.0 per cent of 119 clips were contaminated at the start and end. Settings of 0.75/0.75 s and 0.25/1.00 s were equally clean and 1.00/1.50 s lost 11 clips, so the chosen trim is one of several clean settings, not a unique optimum. The manual audit uses a different sample and criterion, which is why it still records a few leading-edge cases.

Segments that are less than five seconds and those that are longer than thirty seconds are

discarded. The rest are cut by a re-encoding operation that seeks after input decoding, slower than stream copying but placing cut points at the requested timestamps rather than the nearest keyframe, and are written as H.264 at constant rate factor of 23. Each cut is recorded with clip and source details (source URL, raw and trimmed timestamps, duration, selected threshold, cut status); existing clips are skipped unless recutting is requested.

2.6 Handling of Continuous Non-Crash Footage

Non-Crash sources behave differently. Normal driving footage is often uploaded as a single continuous take with no cuts. The boundary procedure then returns one segment covering the whole video, which is longer than the permitted duration and is discarded. Applied unmodified, the segmentation would therefore have produced no negative examples.

These sources were divided into consecutive equal-length intervals within the permitted duration range. This rule generated 371 of the 457 Non-Crash clips. The other 86 came from Non-Crash uploads containing editorial transitions and were segmented with the threshold procedure used for Crash and Near-Miss material. The released index identifies the rule applied to every clip. This practical choice creates a structural difference: most Non-Crash boundaries follow an arbitrary time grid, whereas Crash and Near-Miss boundaries usually follow detected transitions. Boundary structure may therefore correlate with class label and become an unintended visual cue. Section 3.8 tests whether the baselines use this cue.

2.7 Manual Review

Automatic segmentation produced candidates that still required review. The first review tier assigned one of five dispositions - keep, discard, recut, split, or unclear - to 598 candidates. For 467 of these a final decision was made, 358 being retained, 98 removed and 11 set aside for recutting or splitting. The other 131 were left unresolved in this tier; 114 of them, all Near-Miss, entered the corpus through the source-level route described below. In the second tier, 147 of the reviewed clips were subjected to 4 additional structural checks. Reviewers recorded whether the clip had a single scene or multiple incidents, whether either edge of the clip was contaminated and whether the critical event was visible. The other 451 first-tier clips were only labelled with disposition and events.

Candidates that were not part of the structured review were accepted depending on the category of their source upload, with the in-depth review focusing on sources with uncertain categories or questionable segmentation. Thus, the reviewed sample deliberately contains more difficult cases. The final set combines 358 clips accepted after structured review with 963 accepted from source-level categories, producing the 1,321 clips listed in Table 3. Each route demanded a successful cut, duration between five and thirty seconds, and precisely one permitted event label. 358 clips in the structured review were verified by the reviewer as having one continuous scene, with visibility of the event when relevant; those clips that did not qualify were rejected, along with eleven clips that needed recut. No source-admitted 963 clips were individually verified, and so it is not guaranteed that each of them contains precisely one visible event, even though all labels are associated with thematically coherent sources. All 487 Crash clips have been individually verified when we graded their severity. The non-verified clips are in the classes of Non-Crash and Near-Miss. We do not claim that all verified clips are individually verified. The effect is estimated in section 3.4, and the rest of the uncertainty is treated as limitation in section 3.8.

Structural checks give another audit of segmentation. Under-segmented or edge-contaminated clips show those cases when boundary was placed improperly.

2.8 Severity Grading of Crash Clips

Table 2. Criteria for grading apparent physical severity

Grade	Criteria
Major (1). Any one indicator suffices.	Rollover or flip; high-speed or high-impact collision; major deformation or crushing; pile-up of three or more vehicles; vehicle spun or thrown off its path; pedestrian or cyclist struck at speed; debris scattered from the impact.
Minor (0). Impact present but limited.	Low-speed contact or bump; sideswipe, mirror clip or grazing contact; minor denting without structural damage; vehicles remain upright and on their path; single contact producing little or no debris; slow rear-end tap.
Tie-breaking rule	If any Major criterion is clearly visible the clip is labelled Major even if Minor elements are also present. The more severe grade is used if a clip is between the two grades.
Ambiguous cases	If the impact is not fully visible, is occluded, or is cut away before resolution, the annotator records the clip as ambiguous rather than guessing. A clip marked ambiguous by either annotator is excluded from the severity-labelled set. This applies to 54 out of 487 Crash clips.

All Crash-labelled clips were re-examined for severity annotation. All 487 keep their Crash event label and are used in event-level experiments. We established a subsampling plan to measure grading reliability prior to annotation. The full Crash set was graded by the first annotator. For four of the nine Crash-producing uploads, a second annotator independently graded all 297 clips with unlimited replay under identical viewing conditions, using the criteria in Table 2. Agreement from this second pass is reported in Section 3.6. Of the resulting 42 disagreements, a third annotator upheld the first grade in 33 cases and changed it in 9. The adjudicated label file therefore contains 190 Minor and 243 Major clips. For reproducibility, the experiments retain the original benchmark labels of 185 Minor and 248 Major; the corrected file is released as a separate version, and its effect on the Task 3 result is reported in Section 3.6. Severity refers only to physical outcome visible in the footage. The annotator chose ambiguous rather than inferring a grade if the impact was off frame, occluded, or cut before its outcome could be judged. A conservative union rule was used: any clip marked as ambiguous by one annotator was not included in the severity analysis. This was the case with 54 of the 487 Crash clips. They remain ungraded and are not used in the Minor/Major experiments in Section 3.5, leaving 433 graded clips. They are still available as Crash examples for use at the event level. The grading rules are shown in Table 2.

The annotated construct is observable physical severity in video, not injury severity. The result of an injury is dependent on occupant characteristics, use of restraints, vehicle safety systems and medical response. None of these are recorded by a dashcam and the grading does not purport to predict the result. The indicators used are consistent with naturalistic driving research where event severity is based on observable kinematics and physical outcome rather than injury coding [2].

2.9 Source-Aware Partitioning

Clips from the same upload are statistically related. They may share a camera, road environment, watermark, encoding profile or repeated footage of a single incident. Random clip-level splitting could introduce such dependencies on both sides of an evaluation and inflate the reported result. We therefore assign folds by source identifier, not by clip.

There are only 25 uploads, so a single held-out partition would have very few sources and the result would be highly dependent on which uploads were selected. We use four-fold cross-validation grouped by source. Each source is assigned to one of four groups. In each fold one group is used for testing while the other three provide the training data. This process repeats until every clip has been tested once. No source appears in both training and test data within a fold. Validation clips are selected at the clip level from the training sources, so their scores might be overly optimistic; Section 3.8 compares source-grouped validation. Validation is only used to choose the epoch. The test fold remains source-pure, meaning it contains no data from sources used in training. We chose four folds instead of five or ten because only nine sources contribute Crash clips. With more folds, some test partitions would end up with just one or even no source that produces crashes. That would make the severity results unstable or undefined. Four folds ensure that every test partition includes two or three sources that produce crashes. The decision is therefore based on the number of sources, not the number of clips.

2.10 Benchmark Tasks and Metrics

To ensure that later work can compare numbers reliably, GraDE sets up three tasks using the clip set and the same partition. All three tasks follow the four-fold source-grouped protocol described in Section 2.9. Results are macro-averaged over classes and reported as the mean and standard deviation across the four source-pure folds; no single split is used. Accuracy is not the focus because the class proportions in each fold change depending on which uploads are assigned to that fold.

Task 1 - Event Classification: The model assigns each of the 1,321 clips to Non-Crash, Near-Miss, or Crash. Macro-F1 is the primary measure, accompanied by class-wise F1 and a pooled confusion matrix.

Task 2 - Graded Event Classification: Each severity-evaluable clip is assigned to Non-Crash, Near-Miss, Minor, or Major. This task uses 1,267 clips after excluding the 54 Crash clips without a severity grade (Section 2.8). Macro-F1 remains the primary measure. We also report Major recall because a Major-to-Minor error is the more safety-relevant confusion.

Task 3 - Severity Classification: Assign Minor or Major to a clip already known to be a Crash, training and evaluating a two-class model on the 433 severity-graded Crash clips under the same folds. The primary metric is macro-F1 against a chance level of 0.5. This task is defined here because the dataset supports it and because isolating severity from event detection is the natural way to study it. A dedicated two-class CNN + BiLSTM baseline for this task is reported in Section 3.6 (Table 13); the severity figures quoted in Section 3.5 are a looser diagnostic read off the four-class predictions rather than a Task 3 result.

Reference values for Tasks 1 and 2 are given in Section 3.5 and should be treated as numbers to improve on, not as an upper bound. The first Task 3 baseline is reported in Section 3.6 (Table 13).

2.11 Released Artefacts

GraDE is distributed as an annotation index to third-party content rather than as an archival copy, following practices established for other datasets derived from public video platforms [16, 24, 25]. The release comprises a source table giving, per clip, its identifier, source identifier, URL, provisional category, final label, raw and trimmed timestamps, duration and the selected distance threshold with its cosine equivalent; a severity table giving the

Minor or Major grade for each of the 433 graded Crash clips with the *ambiguous* flag for the 54 excluded Crash clips; the source-aware partition assignment; the annotation rubric; and the complete fourteen-threshold per-source composition statistics from which Table 5 is a summary. No video file, extracted frame or derived feature is redistributed. Because the segmentation parameters are reported in full in Sections 2.3 to 2.6 and both raw and trimmed boundaries are released, the clip set can be reconstructed wherever the referenced source version remains available and unaltered; the sixteen-frame banks used below were sampled at uniform positions spanning five to 95% of clip duration and can be regenerated by the same rule. Documentation follows the datasheet convention [24] and the FAIR principles [25].

3. Results and Discussion

3.1 Dataset Composition

Table 3. Distribution of accepted clips. Severity grades cover 433 crash clips (of which 243 non-ambiguous form the agreement sample of section 3.6); 54 ambiguous crash clips are retained as crash examples but carry no grade

Class	Clips	Share of dataset
Non-Crash	457	34.6 %
Near-Miss	377	28.5 %
Crash (all)	487	36.9 %
of which Minor (human-graded)	185	14.0 %
of which Major (human-graded)	248	18.8 %
of which excluded as ambiguous	54	4.1 %
Total	1321	100.0 %

Applying the acceptance rule and source-level admission procedure described in Section 2.7 produced the 1,321 clips that constitute GraDE, drawn from 25 source uploads. Their distribution across the event classes, and the severity grading within the Crash class, are given in Table 3. Consistent with Section 2.8, the Minor and Major totals sum to 433 rather than 487; the 54 clips excluded as ambiguous remain in the dataset as Crash examples. For clarity, the construction funnel is as follows: the pipeline processed 32 source videos, of which 25 contributed clips to the accepted set; these yielded 1,321 accepted clips, of which 487 carry the Crash event label; of those 487, 433 received a Minor or Major severity grade while 54 were excluded as ambiguous; and a predetermined 297-clip subset (reduced to 243 non-ambiguous clips) was double-graded for the agreement analysis of Section 3.6. Figure 2 shows representative frames of the three event classes, and Figure 3 shows examples of the two severity grades.

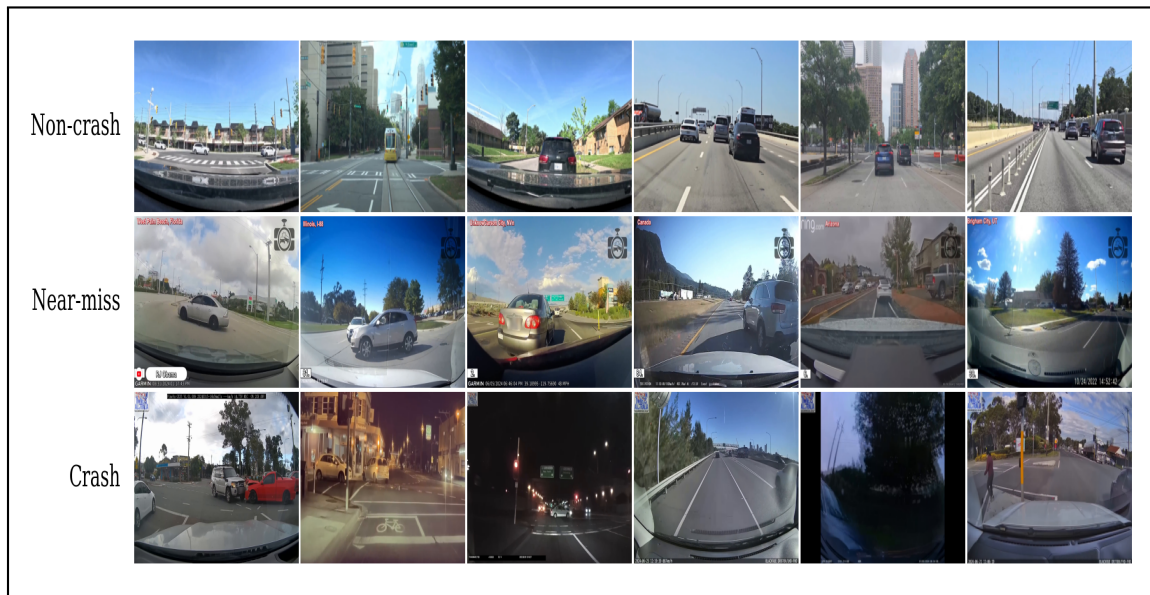


Figure 2. Representative frames from the three event classes (top to bottom: Non-Crash, Near-Miss, Crash), shown after the asymmetric trim of Section 2.5



Figure 3. Representative frames for the two severity grades (top: Minor; bottom: Major), following Table 2. Grades record apparent physical severity, not injury outcome; visible watermarks are the source-level artefacts discussed in Section 3.8

The three event classes are well balanced, none exceeding 37% of the data. Viewed as a four-way problem over Non-Crash, Near-Miss, Minor and Major on the 1,267 severity-evaluable clips, the effective class sizes are 457, 377, 185 and 248, with the smallest holding 14.6%. This is a more even distribution than the three-class view suggests and is favourable for studies using severity as a target.

3.2 Source-Aware Partition

Table 4. Four-fold source-grouped partition. Crash (src) gives Crash clips in test and contributing sources

Fold	Train	Val	Test	Crash (src)
0	852	151	318	115 (2)
1	859	152	310	116 (2)
2	837	148	336	124 (2)
3	820	144	357	132 (3)

Table 4 lists the source-level folds defined in Section 2.9. Among the 25 contributing uploads, 9 supply Crash clips, 8 Near-Miss clips, and 8 Non-Crash clips. Sources are indivisible

and contribute different numbers of clips, so fold sizes are unequal. Test folds contain 310 to 357 clips, including 115 to 132 Crash clips. Their Crash counts sum to all 487 clips (115 + 116 + 124 + 132), with each evaluated once.

Because only nine uploads contain Crash material, each test fold includes two or three such sources and noticeable variation across folds is expected. We consequently report the mean and standard deviation across folds rather than one aggregate score. Source uploads are also thematically homogeneous, and assigning them atomically produces different class proportions in different folds. Macro-averaged class metrics are therefore more informative than overall accuracy. The exact fold assignment is part of the release. The variability visible in Section 3.5 reflects the dataset's limited number of Crash-producing sources, not an unreported change in the evaluation protocol.

3.3 Threshold Selection and Yield

Table 5 contrasts the adaptive choice with fixed thresholds applied to all 32 videos processed by the pipeline; 25 of those videos contributed to the accepted dataset. Except for usable count, the entries are calculated per video and then averaged over the 32 sources. They are not ratios pooled across all generated segments. Usable count is shown only as a corpus total. The same per-video convention is used for the 0.7288 result and for the quantities in the QualityGridAuto objective in Eq. (4).

Table 5. Adaptive selection versus fixed thresholds (representative grid rows). All ratios are means of per-video proportions and sum to one per row; full results for all fourteen thresholds on all thirty-two videos are released as `grade_threshold_grid.csv`

Strategy	Usable (pooled)	Usable ratio (mean of per-video)	Too short	Too long
Fixed $t = 0.20$	1326	0.5947	0.2575	0.1478
Fixed $t = 0.22$	1257	0.6011	0.2152	0.1837
Fixed $t = 0.32$	1042	0.5823	0.0791	0.3386
Fixed $t = 0.46$	456	0.4407	0.0386	0.5207
Adaptive	1134	0.7288	0.0858	0.1854

Adaptive selection chose 12 of the 14 grid values, spanning 0.20 to 0.46; Figure 4 shows the objective for six representative videos. Had one threshold sufficed for this corpus, the selection would have concentrated in a narrow band. On the mean-of-per-video-proportions criterion the usable proportion rises from 0.6011 for the best fixed threshold, $t = 0.22$, to 0.7288 for adaptive selection, a relative improvement of 21.3 per cent computed from unrounded values. The advantage is not driven by outliers: adaptive selection beats the best fixed threshold on 29 of the 32 source videos, ties on two and is lower on one. The lowest grid threshold yields more usable segments in absolute terms only by cutting aggressively, leaving a quarter of its segments below the five-second minimum, and discards 790 unusable segments against 286 for adaptive selection. No tested fixed threshold in the fourteen-point grid attains a better value on all three of usable, over-short and over-long proportion simultaneously.

Threshold and Weight Sensitivity: Across the 14-point grid the mean usable ratio runs from 0.60 at $t = 0.22$ to 0.44 at $t = 0.46$, with the composite objective on a broad plateau between $t = 0.26$ and 0.30 (full 14-point sweep released as `grade_threshold_grid.csv`); the broad plateau and the small differences between neighbouring candidates indicate limited sensitivity within the tested 0.02-spaced grid. No fixed threshold beats 0.60 usable against the adaptive 0.73, and

the maximisers spread over 12 of the 14 values. Table A1 (Appendix A) varies the other settings, using the selection code re-run on the cached similarity scores (it reproduces all 32 selections). A grid step of 0.01 or 0.04, or a range of 0.16 to 0.50, changes the threshold of 4 to 18 videos but keeps the usable ratio within 0.720 to 0.748 and boundary F1 within 0.940 to 0.944; the released grid is kept so that the clip set can be rebuilt exactly. Random weights within ± 50 per cent (2,000 draws) change a median of 3 selections, with usable ratio 0.710 to 0.731. Weights tuned on the manual boundaries (243 vectors, chosen on seven videos and scored on the eighth) give the same held-out F1, 0.944, as the preset weights. The minimum boundary gap is the sensitive setting: 7 s raises the usable ratio to 0.823 but lowers F1 to 0.841 by merging true cuts, and 3 s lowers both. Eq. (4) cannot choose it, so it is fixed at 5 s, the minimum clip duration.

This comparison measures the composition of the candidate segment set, not the accuracy of the detected boundaries. Because usable duration is the dominant term of Eq. (4), the result establishes that per-video selection produces a more usable clip pool for dataset construction, and should not be read as evidence that the boundaries themselves are more accurate. Boundary accuracy against manually marked boundaries is evaluated separately in Section 3.4; the present comparison concerns only candidate-segment composition. Against those manual boundaries, one fixed threshold of 0.22 reaches F1 0.682, compared with 0.944 for per-video selection (Table A1).

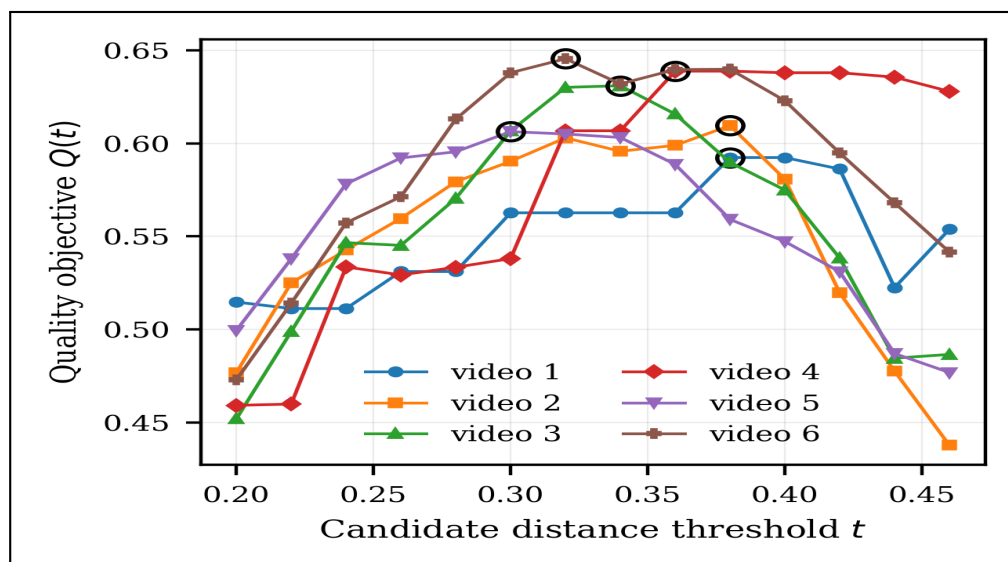


Figure 4. Quality objective $Q(t)$ versus candidate distance threshold for six representative source videos; circles mark the selected maximiser t^* . The maximisers are spread across the grid rather than concentrated

3.4 Segmentation Quality Audit

The structural fields of Section 2.7 audit the segmentation directly, recording how often a human judged a cut clip to be a single coherent event. All four fields were completed for 147 clips. Because the audit sample was deliberately enriched with difficult cases rather than drawn at random, the rates below are descriptive of that sample and should not be generalised to the complete corpus.

Boundary Accuracy Against Manual Marks: To validate the boundaries themselves, not only the segment pool, one annotator marked the true event boundaries in the first four minutes of eight Crash and Near-Miss source videos (126 boundaries), candidate-assisted

(the annotator saw the QualityGridAuto boundaries), so these are candidate-assisted marks rather than fully independent ground truth. Within a 0.5 s tolerance QualityGridAuto reaches precision 0.96, recall 0.93, F1 0.94 (95% CI 0.93 to 0.96, bootstrap over the eight videos) and segment-level temporal IoU 0.93 at a mean absolute offset near zero (temporal IoU is the length-weighted mean, over ground-truth inter-boundary segments, of the best intersection-over-union with any predicted segment), with per-video F1 between 0.92 and 0.97; the nine misses are cuts it did not propose. TransNet V2, a learned shot-boundary detector run on the same videos, scores far lower here (F1 0.54, precision 0.49; F1 difference 0.40, 95% CI 0.27 to 0.49) — not from failing, but because it marks every shot change, including replays and angle switches within a single incident, whereas the boundaries of interest are incident-level. The two agree on single-shot incidents and diverge where an incident spans several shots (Table 14), so a generic shot detector is not a substitute for the incident-aware selection GraDE relies on. TransNet V2's temporal IoU is omitted (Table 14): its over-segmentation fragments the timeline, so segment IoU would reflect fragmentation, not boundary placement.

Of the 147 audited clips, 137 (93.2 per cent) were judged to contain a single continuous scene and 9 (6.1 per cent) were flagged as containing more than one incident, the rate at which the selected threshold under-segmented real material. The remaining clip was recorded as not assessable because the scene structure could not be determined, so the three categories are exhaustive and account for all 147 clips. The critical event was visible in 109 of the 112 clips (97.3 per cent) for which that field was recorded; the field was not completed for the other 35 audited clips. At the edges, the 1.25-second trailing trim eliminated residual content from the following scene entirely, whereas the 0.50-second leading trim left preceding-scene content in 5 clips (3.4 per cent); the asymmetric policy of Section 2.5 is therefore supported in direction, and the leading trim emerges as the tighter of the two constraints.

Non-Crash Label Verification: Non-Crash is a source-level label: clips are extracted from uploads that depict continuous normal driving, so correctness is a property of the source rather than of the individual clip, and the 371 equal-interval Non-Crash clips are correct by construction. The 86 Non-Crash clips drawn from uploads that contain editorial cuts — cut by the threshold procedure rather than the equal-duration grid — are the only subset in which a stray event could be enclosed. A spot check of 50 of these 86 found four clips (8 per cent; 95 per cent Wilson interval 3 to 19 per cent) that in fact contain a near-miss or collision; they are flagged in the released index; the benchmark set retains all 1,321 clips. The implied residual across the whole corpus is about 0.5 per cent and is confined to this event-cut Non-Crash subset. In the Near-Miss class, where per-clip ambiguity genuinely arises, 108 of the 377 clips passed the structured review of Section 2.7 and 269 were admitted from their source category. As a check on the unreviewed clips, the out-of-fold Task 1 predictions agree with the Near-Miss label as often for them as for the reviewed clips (CNN + BiLSTM 0.78 against 0.77; Video Swin-T 0.97 for both; Fisher exact p of at least 0.89). For CNN + BiLSTM, 81 of the 85 disagreements come from two sources in which the reviewed clips disagree as often as the unreviewed ones, which points to a source effect on the model rather than to wrong labels.

3.5 Baseline Benchmarks

As reference points, two baseline systems were trained on GraDE under the four-fold source-grouped protocol of Section 2.9. Both use ResNet-50 initialised with ImageNet weights [22, 23] as the frame backbone with its first six child modules frozen; the two designs differ

only in how per-frame features are aggregated to a clip decision, and share optimiser (AdamW, $\text{lr} = 1\text{e-}4$, batch 8, weight decay $1\text{e-}4$), augmentation (horizontal flip applied consistently across the sixteen frames of a clip) and early stopping (patience 4, at most 15 epochs), and use a cross-entropy loss weighted by inverse class frequency. The three-class models keep the epoch with the best validation macro-F1; the four-class models keep the best three-epoch trailing mean, because their validation folds hold only 16 to 32 clips per severity grade. CNN classifies each of the sixteen sampled frames independently through the ResNet-50 backbone and averages frame logits to produce the clip logit. CNN + BiLSTM extracts a 2048-dimensional descriptor per frame, feeds the sequence through a bidirectional long short-term memory network [26] with hidden size 256, and classifies the time-averaged BiLSTM output. Two dataset views are studied: the three-class event view over Non-Crash / Near-Miss / Crash with 1,321 clips, and the four-class view over Non-Crash / Near-Miss / Minor / Major with 1,267 clips (excluding the 54 crash clips of Section 2.8 that carry no severity grade). Consistent with Section 3.2, all metrics are reported as mean and standard deviation across the four source-pure test folds, macro-averaged over classes. Each configuration is trained with three seeds (42, 1337 and 2024), and the class probabilities of the three runs are averaged within each fold before scoring.

Three-class event baseline. Table 6 reports the event-classification results. Both baselines exceed macro-F1 of 0.86, although part of this score comes from content at the frame corners (Section 3.8). The two cannot be separated statistically. Over the twelve paired fold-by-seed runs, CNN + BiLSTM is ahead by 0.009 on average (paired t-test $p = 0.69$, Wilcoxon $p = 0.47$, ahead in 7 of 12), and a bootstrap over source uploads puts the difference at 0.003 (95% CI -0.037 to 0.055). We therefore claim no advantage for temporal modelling on this task. Per-class F1 shows the Non-Crash and Near-Miss classes are recognised more reliably than the Crash class, whose precision is lower because visually striking non-crash and near-miss clips can share appearance cues with crashes. Table 7 gives the pooled confusion matrix under the CNN + BiLSTM baseline, from which class-level errors can be read directly. A contemporary spatio-temporal model, Video Swin-T [27] pretrained on Kinetics-400 and fine-tuned under the same splits, seeds and selection rule (last two stages trained, mixed precision), reaches 0.9053 ± 0.089 . Its gain comes mainly from Near-Miss (F1 0.955 against 0.852), while its Non-Crash recall is lower (0.80) and its fold scores still range from 0.82 to 0.99, so the event task is not saturated by a stronger video model either.

Table 6. Event three-class baselines. Mean $\pm$ sd across four source-pure test folds; three-seed ensemble

Model	macro-F1	Accuracy	F1 Non-Crash	F1 Near-Miss	F1 Crash
CNN (ResNet-50)	0.8681 ± 0.078	0.8670	0.8974	0.8590	0.8481
CNN + BiLSTM (ResNet-50)	0.8725 ± 0.075	0.8718	0.9116	0.8517	0.8542
Video Swin-T (Kinetics-400)	0.9053 ± 0.089	0.9021	0.8760	0.9552	0.8848

Table 7. Pooled confusion matrix, three-class event baseline (CNN + BiLSTM). Rows are true classes; columns are predicted classes

	pred Non-Crash	pred Near-Miss	pred Crash	total
true Non-Crash	403	1	53	457
true Near-Miss	11	292	74	377
true Crash	12	22	453	487

Flat Four-Class Results: Table 8 evaluates both baselines on the 1,267 clips with severity labels. Relative to their three-class counterparts, macro-F1 decreases from approximately 0.87 to 0.63 for CNN and 0.69 for CNN + BiLSTM. The reduction is concentrated in the severity grades. Non-Crash and Near-Miss remain between 0.77 and 0.93 F1, whereas Minor and Major fall between 0.40 and 0.53. When Minor and Major predictions are collapsed back into Crash, the same outputs give pooled macro-F1 values of 0.8025 and 0.8664 (fold means 0.8012 and 0.8719); Section 3.6 discusses these values. Major recall is also low: 0.520 for CNN and 0.496 for CNN + BiLSTM, and most missed Major clips are predicted as Minor. Fold macro-F1 ranges from 0.50 to 0.74 for CNN and from 0.65 to 0.71 for CNN + BiLSTM, consistent with the two or three Crash-producing sources available in each test fold (Section 3.2). Table 9 contains the pooled confusion matrix, and Table 12 gives precision and recall by class. On this task CNN + BiLSTM is ahead in 9 of 12 paired runs (mean +0.040, paired t-test $p = 0.045$), but the source-level bootstrap interval includes zero (-0.025 to 0.147), so this gap is not established either.

Table 8. Flat four-class baselines on the 1,267 severity-evaluable clips. Mean $\pm$ sd across four source-pure test folds; three-seed ensemble

Model	macro-F1	Accuracy	F1 Non-Crash	F1 Near-Miss	F1 Minor	F1 Major
CNN (ResNet-50)	0.6298 $\pm$ 0.109	0.6747	0.7684	0.8653	0.3997	0.4857
CNN + BiLSTM (ResNet-50)	0.6889 $\pm$ 0.027	0.7579	0.9340	0.8454	0.4511	0.5252

Table 9. Pooled confusion matrix, flat four-class baseline (CNN + BiLSTM). Rows are true classes; columns are predicted classes

	pred Non-Crash	pred Near-Miss	pred Minor	pred Major	total
true Non-Crash	435	1	10	11	457
true Near-Miss	17	307	38	15	377
true Minor	11	15	92	67	185
true Major	12	35	78	123	248

A Conditional Diagnostic on the Severity Cells: Because the four-class confusion matrix mixes two failure modes, it is informative to look at the severity cells alone, restricting attention to clips whose true label is Minor or Major and whose prediction is also Minor or Major. The frame-averaging CNN places 376 of the 433 severity-graded Crash clips in that region and separates Minor from Major within it at macro-F1 0.5779; the CNN + BiLSTM places 360 there and separates them at macro-F1 0.5942. These are not Task 3 results and no two-class severity model was trained. The subset is selected by each model by its own routing, so its size differs (376 against 360), the two numbers are not comparable with each other, and neither is unbiased for the full 433 clips, since a model routing fewer clips into the severity region may be routing the easier ones. It does show that neither system separates Minor from Major far above chance even on clips it has already identified as crashes. Table 10 gives both derived views of the four-class predictions.

Table 10. Derived views of the four-class predictions (not separately trained models). Collapsed rows map Minor and Major back to Crash (pooled macro-F1); severity-only rows score Minor against Major within the severity cells, with n set by each model’s own routing and therefore not comparable across rows. Headline figures are in Tables 6 and 8

Derived view	Model	n	macro-F1
Four-class collapsed to event	CNN	1267	0.8025
Four-class collapsed to event	CNN + BiLSTM	1267	0.8664
Severity cells only (conditional)	CNN	376	0.5779
Severity cells only (conditional)	CNN + BiLSTM	360	0.5942

Relation to Segmentation Quality: Because the classifiers were trained on clips produced by the Section 2 pipeline, their results show that the retained clips contain learnable event information. They are not, however, a direct test of boundary accuracy. A classifier may perform well even when a clip boundary is imperfect, provided that the relevant visual evidence remains. We therefore tested the downstream effect directly. The 17 Crash and Near-Miss sources were segmented twice, with the per-video threshold and with the best fixed threshold (0.22), and every usable segment was labelled by its source class, so neither pool benefits from manual review; the same Non-Crash clips were used in both. A frozen ResNet-50 + BiLSTM trained on each pool (four source-grouped folds, three seeds) and tested on the same released test clips reached macro-F1 0.789 with adaptive and 0.805 with fixed segments, or 0.797 when the fixed pool is cut to the adaptive pool’s size in each class; the fixed pool scored higher in all four folds, by 0.002 to 0.045. Per-video selection therefore does not improve this event classifier, for which a source-level label survives an imperfect cut. Its measured benefits are in segmentation itself, a higher usable proportion and boundary F1 of 0.944 against 0.682 (Table A1), which matter for producing single-incident clips for human grading rather than for coarse event recognition.

Two protocol details should be kept in mind when interpreting the results. First, the 54 Crash clips without a reliable severity grade are included in Task 1 but omitted from Task 2, which explains the different sample sizes. Second, Tables 6 and 8 both average three seeds within each fold, so their standard deviations describe differences between held-out sources; the mean of the twelve single runs is lower (0.8471 and 0.8561 in Table 6, 0.6158 and 0.6556 in Table 8).

3.6 What the Baselines Tell us About Severity Learnability

Taken together, Tables 6-10 support three conclusions about the learnability of severity. These conclusions concern the observed dataset and protocol rather than a claim about a particular architecture; the controlled comparisons retain the same backbone, optimiser, augmentation, and folds. The baselines are reference points that measure how hard the severity decision is on this data; they are not proposed as severity predictors.

Severity is not Merely the Event Task with One Additional Label: Many cues that distinguish Crash from Non-Crash - broken parts, unusual vehicle orientation, or visible deformation - occur in both Minor and Major clips. Separating the two grades depends more on impact magnitude, deformation extent, and post-impact motion, cues that a fixed appearance backbone may represent poorly. The result is a different and substantially harder decision boundary.

Initial Comparison Contains a Confound: Tables 6 and 8 use different clip pools (1,321 versus 1,267), and all 54 clips missing from the four-class pool are Crash clips, the weakest event class. We therefore rescored the three-class predictions on the same 1,267 clips, so that both label spaces share clips, folds, architectures, recipe and seeds.

On these matched clips, macro-F1 falls from 0.8630 to 0.6298 for CNN and from 0.8682 to 0.6889 for CNN + BiLSTM when Crash is divided into Minor and Major, losses of 0.233 and 0.179 that occur in every fold. Collapsing the four-class predictions back to the event classes gives a fold-mean of 0.8719 for CNN + BiLSTM, level with its three-class score, so the severity split costs this model no event accuracy; CNN loses 0.062 (0.8012), mostly in one fold. The reduction therefore comes from the severity grades themselves rather than from weaker event recognition.

Errors in Event Detection do not Fully Account for the Result: Restricting attention to the severity cells gives macro-F1 0.5779 for CNN and 0.5942 for CNN + BiLSTM (Table 10), compared with a chance value of 0.5. These are conditional quantities because each model determines which samples enter its severity cells, so they should not be interpreted as unbiased severity estimates. They nevertheless suggest that poor Minor/Major separation persists after routing. Two factors are plausible: a frozen ImageNet representation may encode appearance better than impact magnitude, and the human grading boundary itself is not perfectly sharp.

A dedicated Minor-versus-Major model, the CNN + BiLSTM of Section 3.5 trained and evaluated only on the 433 graded Crash clips under the same folds, reaches macro-F1 0.5987 ± 0.050 (Table 13) — far below the event scores, with asymmetric errors (Major recall 0.68, Minor 0.53). This is the Task 3 result proper. Class imbalance was addressed directly: inverse-frequency class-weighted and focal ($\gamma = 2$) losses did not help (macro-F1 0.582 and 0.580), shifting recall toward Major rather than lifting Minor — so the difficulty is intrinsic to the severity distinction, not a class-imbalance artefact reweighting can remove. Video Swin-T trained under the same protocol reaches 0.6510 ± 0.008 (Table 13), above CNN + BiLSTM but far below the 0.827 of the second annotator; its single runs average 0.587, so much of the gain comes from the three-seed ensemble.

Table 11. Cross-tabulation of the Minor / Major grade for the 243 non-ambiguous clips graded by both annotators (from the 297 double-graded Crash clips). Rows: first annotator; columns: second

	Second: Minor	Second: Major	Total
First: Minor	96	12	108
First: Major	30	105	135
Total	126	117	243

Human Agreement Contextualises Severity-Task Performance: Both annotators graded the same 297 Crash clips; 54 were marked ambiguous by an annotator and excluded, leaving 243 non-ambiguous clips in the agreement analysis. On these, Cohen’s kappa for the Minor / Major decision is 0.656, with a 95 per cent bootstrap confidence interval of 0.560 to 0.745 and raw agreement of 82.7 per cent; Table 11 gives the cross-tabulation. The boundary is therefore reproducible enough to support a benchmark without being sharp: about one clip in six is graded differently by two people applying the same rubric, and disagreement is asymmetric, the first annotator assigning Major more often than the second. Since the benchmark experiments run on the first annotator’s grades, this disagreement stays represented in the frozen benchmark-label file, while the third-annotator corrections are released separately

in the versioned corrected-label file. This does not set a strict bound on model performance, since a model trained against the first annotator’s labels could agree with that annotator more closely than the second annotator does. It does, however, provide important context for the difficulty of the severity task: a target on which two trained people applying the same written rubric disagree on roughly one clip in six is qualitatively unlike the event classes, which are recognised at macro-F1 above 0.86. A third annotator then adjudicated the 42 disagreements, upholding the first grading on 33 and overturning it on 9 (adjudicated distribution 190 Minor, 243 Major); per-source raw agreement ranges from 0.74 to 0.92. The experiments retain the original benchmark labels (185 Minor, 248 Major) and the adjudicated labels are issued as a versioned update; scored against them, the Task 3 baseline changes by only -0.006 macro-F1 (0.599 to 0.593). As a reference rather than a bound, the second annotator scored against the first reaches macro-F1 0.827, against 0.599 for the Task 3 baseline (0.579 on the same 243 clips), which leaves a wide margin for future models. Disagreement does not explain the low score either. On the 201 clips that both annotators graded alike, the baseline reaches macro-F1 0.56, and its accuracy on the 42 disputed clips (0.71) is not lower than on the agreed clips (0.57; Fisher exact $p = 0.09$), largely because 30 of the disputed clips are Major for the first annotator and the model leans towards Major. The subjective grade therefore adds label noise of about one clip in six, but most of the gap to human agreement lies in the model rather than in the labels.

Statistical Uncertainty: Table A4 (Appendix A) gives 95 per cent confidence intervals for every baseline, from a bootstrap that resamples source uploads within each event class (5,000 draws) and recomputes pooled macro-F1, which differs from the fold mean by at most 0.004. Because the test material comes from only 25 uploads, the intervals are wide, 0.15 to 0.23 for Tasks 1 and 2 and 0.07 to 0.10 for Task 3, and no difference between two models excludes zero. Video Swin-T leads CNN + BiLSTM in 9 of 12 paired Task 1 runs (paired t-test $p = 0.25$, Wilcoxon $p = 0.13$), with fold differences from -0.056 to +0.168, and in all four folds of Task 3 (+0.008 to +0.115; McNemar $p = 0.06$). Model rankings on GraDE should therefore be read as indicative, and comparisons on it should report source-level intervals.

Table 12. Per-class precision, recall and F1 (mean $\pm$ sd across the four source-pure folds) for the flat four-class baselines

System	Class	Precision	Recall	F1
Flat 4-class, CNN	Non-Crash	0.947 $\pm$ 0.029	0.702 $\pm$ 0.321	0.768 $\pm$ 0.226
	Near-Miss	0.878 $\pm$ 0.079	0.861 $\pm$ 0.154	0.865 $\pm$ 0.101
	Minor	0.409 $\pm$ 0.223	0.490 $\pm$ 0.139	0.400 $\pm$ 0.048
	Major	0.453 $\pm$ 0.163	0.534 $\pm$ 0.171	0.486 $\pm$ 0.159
Flat 4-class, CNN+BiLSTM	Non-Crash	0.918 $\pm$ 0.044	0.954 $\pm$ 0.046	0.934 $\pm$ 0.021
	Near-Miss	0.867 $\pm$ 0.111	0.838 $\pm$ 0.185	0.845 $\pm$ 0.132
	Minor	0.468 $\pm$ 0.141	0.502 $\pm$ 0.165	0.451 $\pm$ 0.062
	Major	0.580 $\pm$ 0.094	0.512 $\pm$ 0.147	0.525 $\pm$ 0.076

Table 13. Dedicated Minor-versus-Major severity baseline (CNN + BiLSTM, ResNet-50) on the 433 graded Crash clips (mean $\pm$ sd across the four source-grouped folds; three seeds)

Model	macro-F1	Accuracy	F1 Minor	F1 Major	Recall Minor	Recall Major	n
CNN + BiLSTM (ResNet-50)	0.5987 $\pm$ 0.050	0.6136	0.5330	0.6644	0.5318	0.6814	433
Video Swin-T (Kinetics-400)	0.6510 $\pm$ 0.008	0.6653	0.5871	0.7150	0.5599	0.7507	433

Table 14. Boundary detection vs manually marked event boundaries (8 Crash/Near-Miss videos, first 4 min, 126 boundaries; 0.5 s tolerance)

Method	Precision	Recall	F1	Temporal IoU	Mean abs. offset (s)
QualityGridAuto	0.959	0.929	0.944	0.929	0.00
TransNet V2	0.493	0.603	0.543	—	0.26

3.7 Computational Cost and Scalability

QualityGridAuto costs essentially the same as a single fixed threshold: the threshold search takes about two seconds for the whole corpus on one T4, against an embedding pass of about 1.5 s per minute of source footage, and one four-fold baseline run takes 1 to 1.5 GPU-hours.

3.8 Limitations

Boundary Validation Covers Only Part of the Corpus: Section 3.4 measures precision, recall, F1, and temporal IoU against manual boundaries from eight source videos and includes a TransNet V2 comparison. Equivalent annotation for all source videos has not yet been completed. Because the annotator saw the candidate boundaries, these figures check the rule on a subset and may favour QualityGridAuto; they are not a corpus-level estimate of boundary accuracy.

The controlled Label-Space Experiment Rests on Three Seeds and Four Folds: Section 3.6 compares the same 1,267 clips under identical folds, architectures, recipe and seeds and finds a 0.18 to 0.23 macro-F1 reduction after Crash is separated into Minor and Major; with four source-grouped folds its size is estimated less precisely than its direction.

Annotation Uncertainty was Measured but not Eliminated from the Frozen Benchmark: A third annotator resolved the 42 disagreements found in the reliability subset, changing 9 of 433 grades; the corrected version contains 190 Minor and 243 Major labels. Reported experiments retain the original labels; rescoring the Task 3 predictions against the corrected labels changes macro-F1 by -0.006. On the doubly graded subset, approximately one Crash clip in six received different grades from the first two annotators, with the first annotator choosing Major more often. This estimate has two coverage limitations. The reliability sample represents four of nine Crash-producing uploads, and the event labels were assigned by one annotator without an agreement measurement. Extending double annotation to the other Crash sources and to all three event classes is therefore a priority; an expanded study should again report Cohen's kappa [28].

Fifty-Four Crash Clips have no Severity Grade: In each of these clips, at least one annotator could not see the complete impact and selected ambiguous under the union rule in Section 2.8. The clips remain in GraDE as Crash examples and contribute to Task 1, but they

are excluded from every severity experiment in this paper. Their identifiers and ambiguous status are retained in the release, allowing later users to exclude them or apply a different, explicitly documented review protocol.

Label Verification is Uneven Across the Corpus: Of the 1,321 accepted clips, 358 were checked one by one for scene continuity and event visibility during the structured review. All 487 Crash clips were watched individually during severity grading. The rest of the Non-Crash and Near-Miss clips have labels that came from the original upload. These labels were not checked one by one for each clip. The way the compilation channels are themed supports the class assignment. It does not prove that every single clip has one clear and visible event. Some clips in those two categories may therefore be mislabelled or not properly bounded. The audit of Section 3.4 bounds the rate only on a 147-clip sample enriched for difficult material. The best remedy is to check all 1,321 clips individually for class, single-scene status and event visibility, with a second annotator covering all three event classes. This is the main next step for the event labels, because the agreement study in Section 3.6 concerns severity only.

Source and Selection Bias: Compilation channels select footage for audience interest, which over-represents dramatic events and prevents GraDE from being a probability sample of ordinary driving. Although many Non-Crash clips come from the same channels, and occasionally the same uploads, as the event clips, this also means that all classes share encoding patterns, road environments, and watermarks. A classifier could learn the properties of these sources instead of learning the traffic-event content. Figures 2 and 3 illustrate the risk since visible branding identifies an upload whose content is often thematically consistent. The source-disjoint folds motivate the protocol in Section 2.9 and prevent direct transfer of an upload-specific association from training to test, but they do not make the corpus geographically representative. We do not have data on geographic origin or driving side, so we do not know their distributions, nor do we make any claims that they are balanced. International compilations are likely to contain both left- and right-hand traffic, but either may be correlated with source. Claims of production readiness for autonomous-driving systems would require evaluation on purpose-built, geographically controlled driving data, and are beyond the scope of this work.

The source-bias probe confirms that source information is still visible. A linear classifier identifies the source channel from averaged ResNet-50 descriptors at balanced accuracy 0.333 over 24 sources, compared with chance 0.042. Covering the watermark corners reduces this to 0.311, and the median descriptor cosine remains 0.973. Therefore, the signature is spread over the clip and not just explained by a corner logo, supporting the need for source-separated testing. A test-time masking experiment shows that the Task 1 baselines also draw on the frame periphery (three-seed ensembles, no retraining). Grey boxes over the four corners, each 15 by 25 per cent of the frame, lowered macro-F1 by 0.124 for CNN and 0.077 for CNN + BiLSTM, and a grey 10 per cent border by 0.177 and 0.121, mostly through lower Crash recall (for CNN from 0.92 to 0.66 and 0.55; Table A2). The same four boxes moved to the edge midpoints or to the frame centre changed macro-F1 by only -0.022 to +0.011, so the loss is specific to the corners and border rather than to masking as such. No test upload is seen in training, but several uploads come from the same channel, so the cue can be a channel logo or overlay shared by uploads of one class, or class-typical graphics such as timestamps. This is a form of shortcut learning [29] that source-disjoint testing does not remove, and we recommend reporting a corner-masked

evaluation next to clean results. Downscaling and grayscale cost 0.04 to 0.08.

Heterogeneous Segment Definition: As noted in Section 2.6, 371 Non-Crash clips are bounded by an equal-duration grid, while other clips are delimited by detected transitions. We tested whether the Task 1 baselines use this difference. Each grid clip was re-cut from its source with a duration drawn from the Crash and Near-Miss clips (median 11.3 s instead of 15 s; 5th to 95th percentile 5.2 to 21.3 s), keeping its start inside the same continuous segment, and its 16 frames were re-sampled as before. A control re-cut at the original start and duration reproduced 98 per cent of the original predictions. With event-like durations, Non-Crash recall of the three-seed ensembles stayed at 0.957 for CNN and changed from 0.938 to 0.970 for CNN + BiLSTM, and it was at least 0.92 in every duration band (Table A3), so clip length and frame spacing are not what the models rely on. The 86 Non-Crash clips cut by the threshold procedure have lower recall (0.52 and 0.64) than the grid clips (0.96 and 0.94), but this gap follows the source rather than the cutting route. Of the five sources that contributed clips by both routes, three gave similar recall for both (two at or above 0.95, one at 0.20 for CNN), one gave lower recall for threshold-route clips (0.73 against 0.91 for CNN), and one contributed a single threshold-route clip. The re-cut changes duration but not where a clip starts, so a cue from boundary placement itself is not fully excluded; the released index marks the route of every clip so that the two subsets can be evaluated separately.

Duplicate Incidents: Channels occasionally re-upload the same footage, creating duplicate incidents across uploads, and source-aware partitioning prevents leakage within an upload but not necessarily across uploads. This was checked directly: averaged per-clip ResNet-50 descriptors were compared across sources, and after discounting the generally high similarity of dashcam scenes (median cross-source cosine 0.93), five near-identical pairs stand out at cosine of at least 0.99, all between two Near-Miss sources that the partition had placed in different folds. As these are the only confirmed duplicates, their influence is bounded rather than removed by re-partitioning: re-scoring the two affected folds with the ten duplicated test clips (0.8 per cent of the corpus) excluded shifts macro-F1 by at most 0.002 for every baseline — an order of magnitude below the fold-to-fold standard deviation — so the reported results are unaffected. One clip from each pair is flagged in the released index so users can exclude the redundancy; the benchmark set retains all 1,321 clips.

Validation Leakage: Validation clips are drawn at clip level from the training pool, share source uploads with training and are used to choose the checkpoint. We therefore repeated the CNN + BiLSTM runs (seed 42) with source-grouped validation, holding out in each fold one whole training source per class; the split files are released. The leakage is real for validation scores: clip-level validation overstates test macro-F1 by 0.10 on Task 1 and 0.15 on Task 2, whereas source-grouped validation does not. Its effect on the reported test scores is small once the training data are held constant. Retraining on the full training pool for the number of epochs chosen on the held-out sources changes Task 1 by -0.013 on average (fold range -0.05 to +0.03); on Task 2 it gains 0.02 to 0.06 in three folds, while in the fourth the held-out sources selected a single epoch and macro-F1 fell to 0.38. Training only on the sources that remain after the hold-out lowers Task 1 to 0.768, almost entirely in one fold: without the Non-Crash source S0016, 122 of the 123 Non-Crash test clips, all from S0019, were labelled Crash. With eight Non-Crash sources, removing one changes what the model learns about the class, so the headline tables keep clip-level validation and the source-grouped results are reported as a sensitivity check.

Scale and Source-Diversity Constraint: At 1,321 clips GraDE is small relative to the largest ego-view corpora [10] and is intended to support focused study of graded severity rather than large-scale pretraining. Its 25 source uploads are a stronger constraint than the clip count: clips from one upload are not independent, so the effective sample size for generalisation is closer to the number of sources, and only nine of those contribute Crash clips. This is visible directly in the per-fold spread reported in Section 3.5.

Index-Based Release: Because the release references rather than redistributes the source footage and provides no derived features, reproduction depends on the continued availability of the referenced uploads; clips whose source upload is withdrawn cannot be recovered.

4. Conclusion

This research introduced GraDE, a clip-level dataset of 1,321 ego-view traffic event clips derived from public dashcam compilations and labeled Non-Crash, Near-Miss, and Crash. Of the 487 Crash clips, 433 carry an additional Minor or Major grade of apparent physical severity assigned by human review; the remaining 54 were retained as Crash examples but excluded from severity work as ambiguous. To the extent established here, GraDE is the first released ego-view dataset to pair a near-miss class with graded collision severity for general vehicle-involved events and to distribute a source-disjoint evaluation protocol alongside it. The clips were produced by QualityGridAuto, a construction procedure that replaces the fixed corpus-wide segmentation threshold of prior compilation-derived datasets with a per-video search over a fourteen-point grid. This scores candidates by an objective that rewards usable segment duration, decisive boundaries, and coherent segments while penalizing overly short and overly long ones. It reuses embeddings already computed for boundary detection, adds negligible cost, and raises the mean usable-segment proportion from 0.6011 to 0.7288. On an annotated subset, its boundaries reached F1 0.94 against 0.68 for the best fixed threshold—a gain in clip quality rather than in classifier accuracy. Baselines under a source-pure four-fold protocol reach macro-F1 above 0.86 on the three event classes, partly using content at the frame corners, whereas graded severity is substantially harder. The flat four-class baselines fall to macro-F1 0.63 to 0.69, with the severity classes themselves at F1 0.40 to 0.53, which motivates treating severity as a distinct task rather than an extension of event detection. A third annotator resolved all 42 Minor/Major disagreements in the agreement subset, yielding adjudicated severity labels. Extending double annotation beyond the four double-graded Crash sources and boundary annotation to the full corpus remain the planned extensions.

Declarations

Author Contribution: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing of the original draft, review and editing, and visualization by the first author. Supervision and project administration by the second author. Both authors read and approved the final manuscript.

Funding Statement: None of the authors have received any funding or grants from any institution or funding body for this research.

Conflict of Interest: The authors declare no conflict of interest.

Data and Software Availability Statements

GraDE is distributed as an annotation index to third-party content: the release provides source identifiers, clip timestamps, labels, severity grades, partitions and the annotation rubric, and redistributes no video file, extracted frame or derived feature. The deposited materials are released under CC BY 4.0; this covers those materials only, not the referenced third-party video, which remains subject to the original platform terms and copyright. Users retrieving the uploads are responsible for compliance and must not attempt to identify individuals. The dataset is archived on Zenodo under the DOI 10.5281/zenodo.21787852 and is available at <https://doi.org/10.5281/zenodo.21787852>. Each clip entry carries a SHA-256 checksum of the trimmed clip and of its sixteen-frame descriptor bank, and releases are versioned so that a citation resolves to a specific dataset state. All reported results use the frozen benchmark-label file; a separate corrected-label file flags the four mislabelled Non-Crash clips (Section 3.4) and the five cross-source duplicates (Section 3.8) and contains the nine adjudicated severity grades (Section 3.6).

References

- [1] Park, Jong-Il, Sungyop Kim, and Joon-Ki Kim. "Exploring Spatial Associations between Near-Miss and Police-Reported Crashes: The Heinrich's Law in Traffic Safety." *Transportation research interdisciplinary perspectives* 19 (2023): 100830.
- [2] Dingus, Thomas A., Sheila G. Klauer, Vicki Lewis Neale, Andy Petersen, Suzanne E. Lee, Jeremy Sudweeks, Miguel A. Perez et al. "The 100-Car Naturalistic Driving Study, Phase II-Results of the 100-Car Field Experiment." (2006).
- [3] Bao, Wentao, Qi Yu, and Yu Kong. "Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning." In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, 2682-2690.
- [4] Chan, Fu-Hsiang, Yu-Ting Chen, Yu Xiang, and Min Sun. "Anticipating Accidents in Dashcam Videos." In *Asian conference on computer vision*, Cham: Springer International Publishing, 2016, 136-153.
- [5] Moura, Daniel, Shizhan Zhu, and Orly Zvitia. "Nexar Dashcam Collision Prediction Dataset and Challenge." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 2608-2616.
- [6] Xu, Chuan, Jingqin Gao, Fan Zuo, and Kaan Ozbay. "Estimating Urban Traffic Safety and Analyzing Spatial Patterns through the Integration of City-Wide Near-Miss Data: A New York City Case Study." *Applied Sciences* 14, no. 14 (2024): 6378.
- [7] Jaradat, Shadi, Mohammed Elhenawy, Huthaifa I. Ashqar, Alexander Paz, and Richi Nayak. "Leveraging Deep Learning and Multimodal Large Language Models for Near-Miss Detection Using Crowdsourced Videos." *IEEE Open Journal of the Computer Society* 6 (2025): 223-235.
- [8] Yao, Yu, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. "DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos." *IEEE transactions on pattern analysis and machine intelligence* 45, no. 1 (2022): 444-459.
- [9] Fang, Jianwu, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. "Dada-2000:

- Can Driving Accident Be Predicted by Driver Attention? Analyzed by a Benchmark.” In 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE, 2019, 4303-4309.
- [10] Fang, Jianwu, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. “Abductive Ego-View Accident Video Understanding for Safe Driving Perception.” In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, 22030-22040.
 - [11] Rocky, Arash, Qingming Jonathan Wu, and Wandong Zhang. “Review of Accident Detection Methods Using Dashcam Videos for Autonomous Driving Vehicles.” IEEE Transactions on Intelligent Transportation Systems 25, no. 8 (2024): 8356-8374.
 - [12] Bertasius, G., H. Wang, and L. Torresani. “Is Space-Time Attention All You Need for Video Understanding?” In Proceedings of the International Conference on Machine Learning, 2021, 813–824.
 - [13] Tong, Zhan, Yibing Song, Jue Wang, and Limin Wang. “Videomae: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training.” Advances in neural information processing systems 35 (2022): 10078-10093.
 - [14] Tami, M. A., H. I. Ashqar, M. Elhenawy, et al. “Using Multimodal Large Language Models for Automated Detection of Traffic Safety-Critical Events.” Vehicles 6, no. 3 (2024): 1571–1590.
 - [15] Zhang, R., B. Wang, J. Zhang, Z. Bian, C. Feng, and K. Ozbay. “When Language and Vision Meet Road Safety: Leveraging Multimodal Large Language Models for Video-Based Traffic Accident Analysis.” Accident Analysis & Prevention 219 (2025): 108077.
 - [16] Desai, N. P., A. Etemad, and M. Greenspan. “CycleCrash: A Dataset of Bicycle Collision Videos for Collision Prediction and Analysis.” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2025, 8460–8470.
 - [17] Li, W., H. Xie, and P. Lin. “DSGF-YOLO: A Lightweight Deep Neural Network for Traffic Accident Detection and Severity Classifications.” Vehicles 7, no. 4 (2025): 153.
 - [18] Thounaojam, D. M., K. Rathi, and S. Roy. “Video Shot-Boundary Detection: Issues, Challenges and Solutions.” Artificial Intelligence Review 57 (2024): 104.
 - [19] Souček, T., and J. Lokoč. “TransNet V2: An Effective Deep Network Architecture for Fast Shot Transition Detection.” In Proceedings of the ACM International Conference on Multimedia, 2024, 11218–11221.
 - [20] Santos, K., J. P. Dias, and C. Amado. “A Literature Review of Machine Learning Algorithms for Crash Injury Severity Prediction.” Journal of Safety Research 80 (2022): 254–269.
 - [21] Chen, J., P. Liu, S. Wang, N. Zheng, and X. Guo. “Prediction and Interpretation of Crash Severity Using Machine Learning based on Imbalanced Traffic Crash Data.” Journal of Safety Research 93 (2025): 185–199.
 - [22] Deng, J., W. Dong, R. Socher, et al. “ImageNet: A Large-Scale Hierarchical Image Database.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, 248–255.

- [23] He, K., X. Zhang, S. Ren, et al. “Deep Residual Learning for Image Recognition.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.
- [24] Gebru, T., J. Morgenstern, B. Vecchione, et al. “Datasheets for Datasets.” Communications of the ACM 64, no. 12 (2021): 86–92.
- [25] Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (2016): 160018.
- [26] Hochreiter, S., and J. Schmidhuber. “Long Short-Term Memory.” Neural Computation 9, no. 8 (1997): 1735–1780.
- [27] Liu, Z., J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. “Video Swin Transformer.” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 3202–3211.
- [28] Cohen, J. “A Coefficient of Agreement for Nominal Scales.” Educational and Psychological Measurement 20, no. 1 (1960): 37–46.
- [29] Geirhos, R., J.-H. Jacobsen, C. Michaelis, et al. “Shortcut Learning in Deep Neural Networks.” Nature Machine Intelligence 2, no. 11 (2020): 665–673.

Appendix:

Appendix A. Supplementary Tables

These tables give the full results behind summaries in the main text: the settings sensitivity of Section 3.3 (Table A1), the test-time masking of Section 3.8 (Table A2), the re-cut Non-Crash clips of Section 3.8 (Table A3) and the confidence intervals of Section 3.6 (Table A4).

Table A1. Sensitivity of QualityGridAuto to its settings (usable ratio over 32 videos; boundary F1 over the 8 annotated videos)

Setting	Videos changed	Usable ratio	Distinct thresholds	Boundary F1
Default (step 0.02, gap 5 s)	–	0.7288	12	0.944
Grid step 0.01	12	0.7315	18	0.944
Grid step 0.04	18	0.7204	7	0.940
Range 0.16–0.50	4	0.7480	14	0.944
Minimum gap 3 s	13	0.6879	11	0.872
Minimum gap 7 s	13	0.8226	12	0.841
Random weights $\pm 50\%$	median 3	0.7095–0.7310	–	–
Weights tuned on boundaries	–	–	–	0.944
Fixed $t = 0.22$	–	0.6011	1	0.682

Table A2. Task 1 macro-F1 under test-time masking (three-seed ensembles, no retraining; change from clean in brackets) and Crash recall. Boxes are 15×25 per cent of the frame; the border is 10 per cent

Test-time condition	CNN	CNN + BiLSTM	Crash recall (CNN / BiLSTM)
Clean	0.868	0.873	0.92 / 0.93
Grey boxes at the four corners	0.744 (-0.124)	0.796 (-0.077)	0.66 / 0.85
Same boxes at the edge midpoints	0.873 (+0.005)	0.884 (+0.011)	0.92 / 0.94
Same boxes around the centre	0.846 (-0.022)	0.868 (-0.005)	0.92 / 0.93
Grey 10 per cent border	0.691 (-0.177)	0.753 (-0.121)	0.55 / 0.80
Central 80 per cent crop	0.706 (-0.162)	0.737 (-0.136)	0.69 / 0.85
Downscale to 112 px and back	0.830 (-0.038)	0.791 (-0.083)	0.93 / 0.94
Grayscale	0.810 (-0.059)	0.792 (-0.082)	0.90 / 0.89
All four transforms together	0.291 (-0.577)	0.241 (-0.632)	0.99 / 1.00

Table A3. Non-Crash recall of the 371 grid-cut Non-Crash clips after re-cutting to event-like durations (three-seed ensembles of the Task 1 models, no retraining)

Clips	n	CNN	CNN + BiLSTM
Original clips	371	0.957	0.938
Control re-cut (original start and duration)	371	0.951	0.957
Re-cut, all durations	371	0.957	0.970
Re-cut, 5 to 8 s	61	0.918	0.967
Re-cut, 8 to 12 s	127	0.984	0.984
Re-cut, 12 to 16 s	95	0.937	0.958
Re-cut, 16 to 20 s	58	0.966	0.948
Re-cut, 20 to 30 s	30	0.967	1.000

Table A4. Pooled macro-F1 with 95% confidence intervals (bootstrap over source uploads within each event class, 5,000 draws) and differences between baselines

Task	Model or difference	Pooled macro-F1	95% CI
1 (three-class)	CNN	0.8655	0.723 to 0.952
1 (three-class)	CNN + BiLSTM	0.8686	0.738 to 0.951
1 (three-class)	Video Swin-T	0.9066	0.819 to 0.970
2 (four-class)	CNN	0.6311	0.522 to 0.722
2 (four-class)	CNN + BiLSTM	0.6889	0.600 to 0.746
3 (Minor/Major)	CNN + BiLSTM	0.6025	0.556 to 0.660
3 (Minor/Major)	Video Swin-T	0.6533	0.611 to 0.677
1	CNN + BiLSTM minus CNN	+0.003	-0.037 to 0.055
1	Video Swin-T minus CNN + BiLSTM	+0.038	-0.034 to 0.152
2	CNN + BiLSTM minus CNN	+0.058	-0.025 to 0.147
3	Video Swin-T minus CNN + BiLSTM	+0.051	-0.030 to 0.105