A. Additional Dataset Statistics

The ACCIDENT includes 2,027 surveillance clips whose durations range from 1 to 114 seconds (median: 26.8s), having a median of 375 frames. Frame rates range from 4 to 50 FPS, reflecting the heterogeneity of online CCTV sources. Resolution ranges from 314p to 3840p, but visual clarity is often compromised by heavy compression, motion blur, and poor lighting. Therefore, around two-thirds are of poor or very poor quality.

Environmental diversity is relatively wide. Despite the majority of the videos being in normal weather (81.5%), there are relatively large numbers of videos in rain (13%) or snow (5.5%), which add valuable edge cases. Nighttime scenes account for about a third of the videos, enabling evaluation under low-visibility conditions. Scene types are varied: almost half (46.2%) occur at highways, with the remainder split across intersections, and other traffic layouts.

Accident types are distributed relatively evenly across the five defined structural types (e.g., head-on, sideswipe), ensuring broad task coverage and, importantly, reflecting the real distribution. Spatial annotations typically cover 0.7% of the frame area, with most bounding regions falling within the 0.1–4.8% range.

B. Licensing & Data Availability

To ensure lawful and ethical use and reproducible evaluation, we define the licensing policy, describe the artifacts released, and establish privacy safeguards for ACCIDENT.

Policy & provenance. Only clips whose upstream licenses *explicitly permit redistribution and derivative works* (e.g., CC-BY, open-government reuse terms) were included; all others were excluded. For each released clip, the metadata stores a unique identifier and the canonical source URL.

Privacy. Given the overall small visual quality and the apparent size of scene agents, no clip contains visible faces; license plates may appear incidentally. No identity information is collected or annotated. Use is restricted to research, and any attempt at re-identification is prohibited. A takedown channel will be provided for removal requests.

Released artifacts & licenses. We release: (i) downloadable videos for all licensed clips; (ii) per-clip annotations (impact time and location, and collision type) and metadata; (iii) the evaluation toolkit; and (iv) CARLA assets/code for development. Annotations/metadata are under `CC BY 4.0`; code and CARLA assets are under `Apache-2.0`.

Hosting. The ACCIDENT benchmark website, evaluation toolkit, code, and dataset access instructions are publicly available. The code is hosted on GitHub, and the dataset is distributed through Kaggle, and the leaderboards are available on the custom Website.

C. Inter-Annotator Agreement (IAA)

In order to quantify annotators' uncertainty and the reliability of the ground truth, we provide an inter-annotator agreement (IAA) for all three tasks. These measurements establish an upper bound and allow us to better define σ_t, σ_x, and σ_y. Below, we report the per-annotator accuracy for six annotators[5], and *mean* for all. Besides, we provide confusion matrices for collision types for the best and worst annotators to expose systematic/semantically local confusions (see Fig. 5).

Results. Table 6 shows per-annotator accuracies for those with >700 labeled clips. Agreement is highest for *spatial localization* (mean 0.995), followed by *impact frame* (mean 0.979). *Collision type* is lower on average (mean 0.923) and exhibits larger cross-annotator variance ($\pm$0.026), reflecting semantic ambiguity between certain classes (e.g., t-bone vs. sideswipe). These patterns are also clearly visible in the confusion matrices for A2 and A3 (Fig. 5).

Implications. The relatively larger spread for *collision type* motivates clearer taxonomy definitions and examples in the guidelines. High *spatial* and *temporal* agreement under our tolerance protocol supports the use of Gaussian similarity for evaluation (Sec. 5), which softly penalizes small deviations while avoiding brittle thresholds.

	A1	A2	A3	A4	A5	A6	mean $\pm$std
$\mathcal{T}$	0.978	0.982	0.986	0.979	0.979	0.973	0.979$\pm$0.004
$\mathcal{S}$	0.993	0.996	0.997	0.995	0.996	0.993	0.995$\pm$0.001
$\mathcal{C}$	0.949	0.870	0.915	0.931	0.928	0.944	0.923$\pm$0.026

Table 6. **Inter-annotator agreement accuracy.** Per-task accuracies against the consensus. Annotators with >700 clips only. Rows denote $\mathcal{T}$=*temporal*, $\mathcal{S}$=*spatial*, $\mathcal{C}$=*collision type*. For temporal localization, σ_t=1 is used.

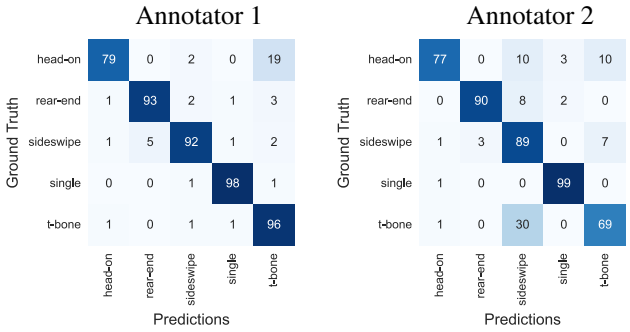

Figure 5. **Collision-type annotation confusion.** As expected, *errors* concentrate between semantical similar classes (e.g., *t-bone* vs. *sideswipe*), with just handful of unrelated confusions.

[5]The A6 was the least active annotator with 700 annotated videos.