# SynSHRP2: A Synthetic Multimodal Benchmark for Driving Safety-critical Events Derived from Real-world Driving Data

Anonymous CVPRW SynData4CV submission

Paper ID 32

## Abstract

*Driving-related safety-critical events (SCEs), including crashes and near-crashes, provide essential insights for the development and safety evaluation of automated driving systems. However, two major challenges limit their accessibility: the rarity of SCEs and the presence of sensitive privacy information in the data. The Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS)—the largest NDS to date—collected millions of hours of multimodal, high-resolution, high-frequency driving data from thousands of participants, capturing thousands of SCEs. While this dataset is invaluable for safety research, privacy concerns and data use restrictions significantly limit public access to the raw data. To address these challenges, we introduce SynSHRP2, a publicly available, synthetic, multimodal driving dataset containing over 1,874 crashes and 6,924 near-crashes derived from the SHRP 2 NDS. The dataset features de-identified keyframes generated using Stable Diffusion and ControlNet, ensuring the preservation of critical safety-related information while eliminating personally identifiable data. Additionally, SynSHRP2 includes detailed annotations on SCE type, environmental and traffic conditions, and time-series kinematic data spanning 5 seconds before and during each event. Synchronized keyframes and narrative descriptions further enhance its usability. This paper presents two benchmarks for event attribute classification and scene understanding, demonstrating the potential applications of SynSHRP2 in advancing safety research and automated driving system development.*

## 1. Introduction

Driving safety research is increasingly using approaches based on artificial intelligence to tackle tasks such as crash detection, scene understanding, driver monitoring, and unsafe maneuver detection. These advancements rely on high-quality multimodal datasets with accurately labeled safety-critical events (SCEs), including crashes and near-crashes. However, the rarity of such events, high collection costs, lack of data labels, and strict privacy regulations make accessing suitable datasets challenging, hindering progress and fair benchmarking across models.

A common limitation for publicly accessible transportation safety data has been a lack of SCEs due their rarity [5, 18, 24, 65]. The average police-reported crash rate in the US is 3.29 crashes per million miles traveled in 2022, and thousands of hours of data collection are needed to capture a single crash [44]. The majority of the publicly accessible data is not large enough to capture a reasonable number of crashes for robust analysis. For instance, the popular BDD100k dataset contains over 1,000 hours of driving videos with time-series data but lacks labeled SCEs [65].

Large-scale naturalistic driving study (NDS) could be a valuable source to address the rarity of SCEs. NDS is characterized by continuous driving data collection using multiple sensors instrumented on participants' vehicles under natural driving conditions. The largest NDS to date, the Second Strategic Highway Research Program (SHRP 2) contains millions of hours of driving data with thousands of SCEs identified [14]. The SHRP 2 NDS dataset required investing approximately $155 million in data collection, storage, and management, resulting in over 1,000,000 hours of driving data. The collected data include four camera views, 3-D acceleration, radar, yaw rate, GPS, and lighting [17].

High-quality annotation is another major challenge. Driving scenarios are complex and involve environmental factors, traffic flow conditions, traffic control, the ego vehicle, and driver behavior. Annotating and labeling SCEs requires expertise of the safety domain and considerable resources. For example, the SHRP 2 NDS undertook a comprehensive research project to develop effective annotation methodologies for its extensive dataset. Factors directly related to driving and how to operationally annotate a driving scene is a huge undertaking [22].

However, due to strict privacy policies, access to such data requires significant effort and costs [38]. As with SHRP 2, researchers must undergo rigorous user certifi-

cation processes, including Institutional Review Board approval, and are granted only time-limited access under strict data use policies.

Artificially generated synthetic data, produced by platforms such as Wayve's GAIA-1 [23] and the CARLA simulator [15], offers an alternative source of multimodal driving data. However, challenges persist with synthetic data in SCE-related research: accurately configuring the numerous parameters influencing SCEs is complex; ensuring high fidelity to real-world conditions is difficult; and synthetic datasets may not fully capture the variability of actual driving scenarios, limiting their practical applicability.

Another challenge in publicly accessible multimodal driving data is the lack of benchmarking [2, 4, 7, 28, 35, 37, 40, 46, 49, 52, 52, 56, 62]. For instance, both Shi et al. [46] and Arvin et al. [2] utilize the SHRP 2 NDS dataset for crash detection algorithm development but with different configurations. Shi et al. [46] uses 59,997 normal driving instances, 1,820 crashes, and 6,848 near-crashes for three-way classification (crash vs. near-crash vs. normal driving), whereas Arvin et al. [2] selects a smaller subset consisting of 7,566 normal driving instances and 1,315 crashes and near-crashes for two-way classification (crash/near-crash vs. normal driving). This variability in data usage, combined with differences in implementation and hyperparameter tuning, makes it challenging to compare results fairly. These challenges highlight the urgent need for standardized datasets and evaluation protocols to enable consistent and fair benchmarking in driving safety studies.

This study advances driving safety assessment by introducing a fully public driving safety evaluation dataset and establishing a benchmark for SCE risk evaluation, including attribute detection and scene understanding. The dataset integrates multimodal data—tabular records, time-series signals, keyframe images, and natural language descriptions—to support research in crash prediction, driving behavior analysis, and multimodal learning. To balance privacy and data utility, we develop a Stable Diffusion-based workflow with ControlNet to de-identify personally identifiable information (PII) while preserving critical driving context. Derived from the SHRP 2 NDS, the dataset contains 1,874 crashes and 6,924 near-crashes, each labeled with event type, conflict type, and incident type. It also includes five key timestamps, time-series sensor data spanning 5 seconds before and during each event, and annotated narrative descriptions, providing a comprehensive resource for benchmarking event attribute classification and scene understanding in driving safety research.

The rest of the paper is organized as follows: related works are discussed in Section 2; Section 3 and 4 detail the SynSHRP2 data and processing workflow; Section 5 presents two tasks along with their corresponding benchmarks; and summary and conclusion are provided in Section 6.

## 2. Related Works

### 2.1. Publicly accessible multimodal driving datasets

There are two main classes of multimodal driving datasets: real-world and synthetic. Real-world datasets include nuScenes, which offers a comprehensive, multimodal sensor suite for complex urban scenarios [5]; the Waymo Open Dataset, which provides synchronized multi-sensor data—incorporating LiDAR, radar, and multiple cameras—along with detailed 3D object annotations [51]; and KITTI, a benchmark that has set standards for object detection and tracking [18]. Naturalistic datasets such as SHRP 2 capture high-frequency video and vehicle telemetry from thousands of drivers in real-world settings, offering detailed insights into driver behavior, distraction, and crash-risk factors that are critical for enhancing road safety [22]. BDD100K offers a large-scale collection of driving videos and annotated images across various weather and lighting conditions [65]; Brain4cars focuses on driver maneuver anticipation using both in-cabin and exterior views [24].

On the synthetic side, WayveScenes101 is a high-resolution dataset designed for novel view synthesis in autonomous driving [70]. Several CARLA-derived datasets are also available. These include KITTI-CARLA [12] for real-to-synthetic comparisons, CarlaSC [55] for semantic scene completion, CARLA-Loc [21] for simultaneous localization and mapping evaluation, and Paris-CARLA-3D [13] for dense point clouds. They offer a diverse suite of synthetic resources that enable robust scene reconstruction and performance evaluation under various simulation scenarios.

While these datasets offer valuable insights into autonomous driving and driving safety, research on SCEs is constrained by data limitations. Real-world datasets struggle to capture SCEs due to their rarity and privacy concerns, while synthetic datasets cannot yet fully replicate the complexity and nuance of SCEs.

### 2.2. Stable Diffusion models for synthetic generation

Stable Diffusion [42] is a class of diffusion models designed for efficient computation by operating on latent representations extracted through an autoencoder [19]. This approach significantly reduces the computational cost while maintaining high-quality image synthesis. Stable Diffusion is one of the state-of-the-art methods for generating synthetic images and has become a powerful tool for realistic image generation [10], upscaling [34], image denoising [27, 69], and video generation [58]. Its versatility enables applications across diverse fields, including graphic design [66], animation [45, 60], music production [50], and robotics

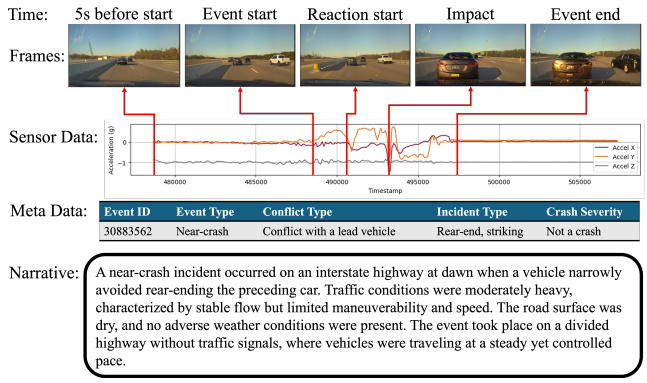


Figure 1. Example illustrating all data types in SynSHRP2.

[6, 30].

ControlNet [67] is a trainable neural network structure designed to guide the generation process of a pretrained Stable Diffusion model. It enables users to impose additional conditions, such as Canny edges, human pose, and text prompts, ensuring that the generated images adhere to specified constraints. This flexibility extends the applicability of ControlNet to image editing and manipulation [59], including artistic style transfer [61], object composition [29], reconstruction and restoration [32], and medical imaging [26]. A notable extension, IP-Adapter [64], refines ControlNet by enhancing image prompt capabilities, ensuring that generated images closely resemble the input reference image.

## 3. SynSHRP2 Dataset

The SynSHRP2 dataset is a high-quality, synthetic, multimodal dataset of real-world SCEs designed to advance research in driving safety, vision-language model (VLM) development, and automated driving system (ADS)/advanced driver assistance system (ADAS) evaluation. SynSHRP2 provides multimodal data on SCEs, including time-series kinematic signals, synthetic images of SCE scenarios, detailed annotations, and event narrative descriptions. SynSHRP2 utilizes Stable Diffusion with ControlNet to accurately de-identify PII while preserving critical safety-related information, ensuring privacy protection without compromising data integrity. This section introduces the detailed dataset setups and the methodology for de-identified synthetic scene generation workflows.

### 3.1. Dataset setups

The dataset consists of 1,874 crashes and 6,924 near-crash events, organized into four components: tabular records, sensor data, keyframe images, and comprehensive narrative descriptions of events. An example illustrating all modalities is shown in Figure 1.

**Tabular records**. The tabular records provide detailed annotations for each SCE, capturing essential information about the event's context and severity. Key fields include *Event ID* (unique identifier), *Timestamps of Keyframes* (capturing five critical moments: 5 seconds before Event Start, Event Start, Reaction Start, Impact, and Event End), and *Event Type* (classifying events as crash or near-crash). *Conflict Type* identifies the objects involved in the conflict (e.g., lead vehicle, following vehicle, and parked vehicle), with multiple conflicts listed in sequence, prioritizing the most severe. *Incident Type* specifies the nature of the conflict (e.g., rear-end collision, road departure), while *Crash Severity* ranks the crash event based on vehicle dynamics, property damage, known injuries, and risk level to drivers and road users. SynSHRP2 includes two event types, 16 conflict types, 18 incident types, and four levels of crash severity, offering comprehensive description for understanding the diverse nature of SCEs. Detailed descriptions of these fields are provided in the appendix.

**Sensor data**. The sensor data, recorded using an inertial measurement unit (IMU) and radar sensors, provides detailed time-series measurements, spanning from 5 seconds before the Event Start to 5 seconds after the Event End. Key fields include *Timestamps* for precise temporal alignment, *Longitudinal, Lateral, and Vertical Accelerations*, and *Speed* to track vehicle dynamics. *Pedal Brake State* indicates braking activity, while *Lane Width*, *Left Line Right Distance*, and *Right Line Left Distance* capture the vehicle's lane position. This continuous sensor data offers a dynamic view of the vehicle's behavior, complementing the tabular records and providing comprehensive description of driver response and vehicle control during SCEs. Detailed descriptions of these sensor data fields are provided in the appendix.

**Event narratives**. Narrative descriptions are provided for each SCE, manually annotated by trained data coders. These annotations capture key contextual details such as traffic density, lighting conditions, road surface conditions, locality, event type, conflict type, and incident type, offering rich insights into each SCE.

**Synthetic Keyframe images**. The keyframes correspond to five critical timestamps: 5 seconds before Event Start, Event Start, Reaction Start, Impact, and Event End, each image with a resolution of 1920 × 1080. These keyframes are extracted from SHRP 2 NDS front-view videos and have undergone a PII de-identification process, detailed in Section 3.2.

### 3.2. Synthesize De-identified Keyframes

This process synthesizes de-identified keyframe images from the SHRP 2 NDS front-view video dataset with the following objectives: 1) Upscale resolution; 2) Protect PII, including vehicle stickers, street names, and pedestrians;

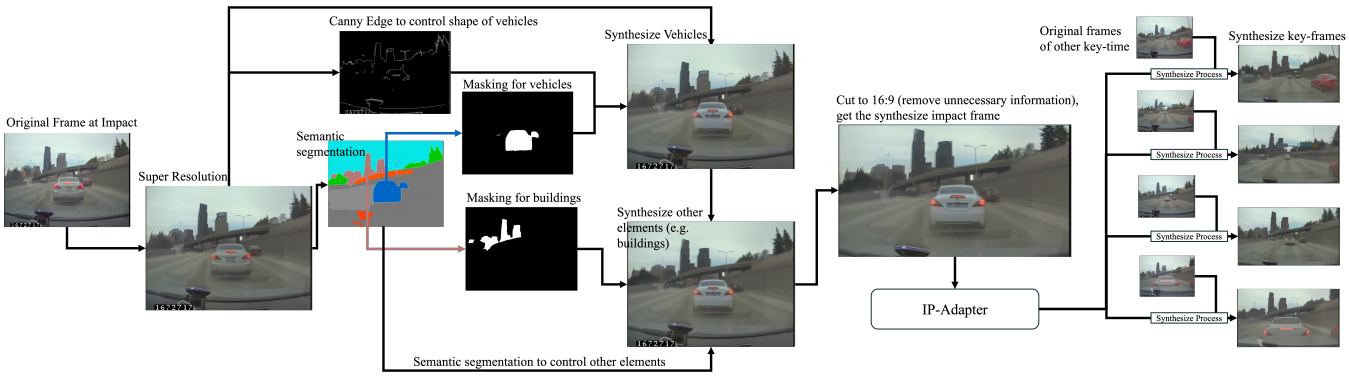


Figure 2. Workflow of the de-identified keyframe synthesis process. The pipeline begins with a keyframe input, applying super-resolution to upscale image resolution. Semantic segmentation then identifies object-related pixels, while Canny edge detection is used for objects where orientation is critical. The new frame is synthesized part by part by masking corresponding segments. Image cropping removes unnecessary elements like timestamps. For the "Impact" keyframe, this completes the process. For subsequent keyframes, an IP-adapter utilizes the first keyframe as an image prompt to ensure consistency across frames.

3) Preserve essential traffic information, such as spatial-temporal relationships among road users, key traffic scene setups (e.g., intersections, highways, rural roads), and traffic control signs/devices; 4) Ensure consistency across consecutive frames; and 5) Remove irrelevant elements like video timestamps and the vehicle's front hood.

To achieve this, we developed a comprehensive video frame synthesis method consisting of five key components: 1) StableSR [54] is applied to the selected keyframes of the videos in the SHRP 2 NDS dataset for image upscaling. 2) Semantic segmentation [20] is applied to the upscaled frames to achieve pixel-level precision in generalization control. 3) Through stable diffusion with semantic segmentation and Canny Edge Detection ControlNets [43], detected objects are reproduced on the corresponding segments. For objects whose orientation can contain important traffic information, line sketches are used to control the orientation of the reproduction. 4) IP-adapter [64] is applied on "Component 3" to ensure the consistency of the generated objects in each frame of the videos. 5) Image cropping is applied to remove timestamps and the vehicle's front hood. This method can be extended to synthesize other de-identified datasets. Figure 2 illustrates the workflow of the proposed approach. The following sections describe the first four components when applied to keyframes from a general video.

**Stable Diffusion-based upscaling**. We use StableSR (**SR**) [54] for upscaling the keyframes. Specifically, denote $I_t$ as the $t$-th frame of the video with size $N \times N$, then the corresponding upscaled image can be denoted as

$$\hat{I}_t = \mathbf{SR}(I_t) \tag{1}$$

where $\hat{I}_t$ is an image with size $M \times M$, $M > N$.

**Sematic segmentation for pixel-level object classification**. To identify objects with PII (e.g., vehicles and pedestrian) in a keyframe, we use semantic segmentation (**Seg**) to classify the pixels therein. The output of the process can be written as

$$\{\mathcal{P}_1, \cdots, \mathcal{P}_O\} = \mathbf{Seg}(\hat{I}_t) \tag{2}$$

where $\mathcal{P}_O$ is the segment of pixels classified as object $O$.

**Synthesis of de-identified keyframes**. To protect PII while retaining critical traffic information, a keyframe is reproduced part by part based on the classes of objects detected by semantic segmentation. Denote $\tilde{I}_t = \{\tilde{\mathcal{P}}_1, \cdots, \tilde{\mathcal{P}}_O\}$ as the synthesized frame with semantic segmentation based on $\hat{I}_t$, where $\tilde{\mathcal{P}}_O$ is the synthesized segment corresponding to segment $\mathcal{P}_O$[1]. For each object $O$, if object $O$ contains no PII, the corresponding segment is directly passed to the blue synthesized frame, i.e., $\tilde{\mathcal{P}}_O = \mathcal{P}_O$. If object $O$ contains PII, its corresponding segment $\tilde{\mathcal{P}}_O$ will be synthesized through a large pre-trained stable diffusion model (**SD**) [42] with ControlNets (**ConNet**) [67] via the following process to achieve de-identification:

1. To retain critical traffic information, the synthesized segment must maintain the same orientation as the original. For instance, if the original segment depicts the rear of a vehicle, the generated segment should also show the rear, not the front, ensuring the accuracy of the traffic scenario. To achieve this, we extract the outline sketch of the segment $L_O$ using the Canny Edge Detection algorithm (**Canny**) [43]: $L_O = \mathbf{Canny}(\mathcal{P}_O)$.
2. To control the properties of the segment to be synthesized, we use text prompts $T_O$, consisting of positive prompts $PT_O$ and negative prompts $NT_O$, i.e., $T_O = \{PT_O, NT_O\}$. Positive prompts indicate desired prop-

[1]Segment $\tilde{\mathcal{P}}_O$ and segment $\mathcal{P}_O$ should have the same number of pixels for all obejcts $O$.

erties, such as "high quality," "naturalistic," and the object's semantic meaning ($O$). Negative prompts specify undesired properties, including "low quality," "cartoon," and "watermarks."

3. To incorporate the above information into the stable diffusion process for synthesis of the segment, two ControlNets are added on top of a pre-trained large diffusion model. One ControlNet's inputs are the outline sketch $L_O$ and the text prompts $T_O$. This ControlNet controls the diffusion model of the properties and orientation of the synthesized segment. The other ControlNet's inputs are the original segment of the object $\mathcal{P}_O$. This ControlNet guides the diffusion model in determining the segment's location within the image. Then, the segment $\tilde{\mathcal{P}}_O$ is reproduced as

$$\tilde{\mathcal{P}}_O = \textbf{SD}(\textbf{ConNet}(L_O, T_O), \textbf{ConNet}(\mathcal{P}_O); z), \quad (3)$$

where $z \sim \mathcal{N}(0, 1)$ is used for sampling images, enabling image generation based on the diffusion model's density function.

**Consistent object representation with IP-Adapter**. To ensure consistent object representation across consecutive keyframes, we adopt IP-Adapter [64]. The previously synthesized frame serves as the image prompt to guide the diffusion process. Specifically, **IP-Adapter**$(\tilde{I}_{t_{Impact}})$ functions as an additional ControlNet alongside the two existing ControlNets in Equation (3) to facilitate the generation of keyframes $I_t$, for $t = 1, 2, \cdots$, assuming $I_0$ is the first keyframe. The generation of $I_0$ follows Equation (3) directly.

# 4. Implementation of Synthetic Image Generation

The platform is ComfyUI v0.3.14 [8], running on Python 3.10 and Rocky Linux 9.3, with model training performed on a workstation equipped with dual Intel Xeon Gold 6338 CPUs, 256 GB RAM, and two Nvidia Tesla A100 (80 GB) GPUs. For super-resolution, we employed StableSR [54] with its ComfyUI node implementation [57]. The de-identified keyframe synthesis was performed using Stable Diffusion XL (SDXL) [39] as the base model, leveraging the RealArchVisXL checkpoint [25]. The Canny and Segmentation ControlNet modules were implemented via ComfyUI-ControlNet-Aux [16], while ComfyUI-IPAdapter-Plus [11] was used for image adaptation.

## 4.1. Module effectiveness

This section compares the proposed approach with alternative methods, including Upscale + Masking, Canny ControlNet, and IP-Adapter ControlNet, to demonstrate its effectiveness in synthetic image generation and privacy de-identification.

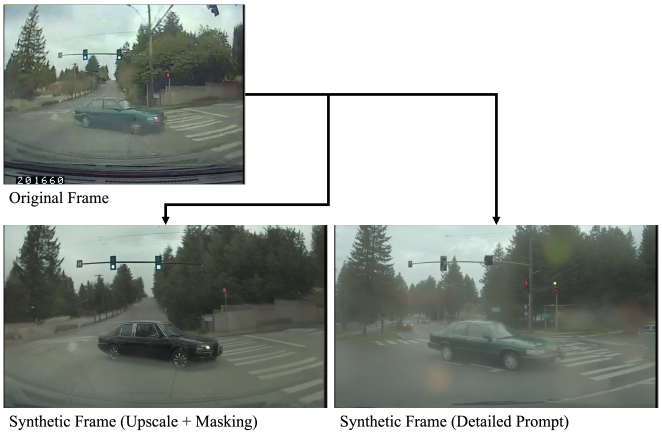


Figure 3. Comparison of synthetic image generation approaches.

**Upscale + Masking process**. The proposed resolution upscale + semantic segmentation ControlNet approach is designed to maintain the structural integrity of critical driving safety-related elements while ensuring effective de-identification. An alternative method is img2img generation with a detailed text prompt. As shown in Figure 3, a synthetic frame was generated using the alternative approach with CFG = 3.5, denoise = 0.4, and other hyperparameters set identically. The prompt used was:

"A suburban intersection on an overcast day, viewed from a dashcam perspective. An older four-door sedan is turning left through the intersection under a green traffic light. Tall evergreen trees line the background, and there's a sidewalk on the right side. The scene is photorealistic with natural, muted daytime lighting, capturing the sense of a real-life moment in motion."

The Upscale + Masking approach preserves key driving safety-related infrastructure, such as roads, traffic lights, and pavements, which appear in the same or highly similar patterns as the original frame. The car models and background greenery are modified for de-identification. However, in the detailed text prompt approach, despite explicitly mentioning details like "intersection under a green traffic light" and "a sidewalk on the right side," the generated frame fails to maintain spatial consistency. The road direction changes, an additional vehicle appears that was not present in the original frame, and traffic lights are misplaced with incorrect colors, distorting SCE evaluation. This comparison highlights the advantages of the upscale + masking process method in maintaining structural consistency while ensuring privacy preservation.

**Canny ControlNet**. If semantic segmentation is used as the sole control mechanism, it presents a limitation in accurately preserving the directionality of vehicles, such as distinguishing between the front and rear. As shown in Figure 4, the highlighted car in the original frame represents the

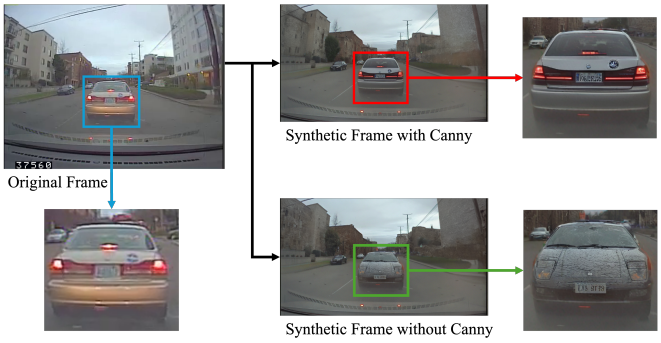


Figure 4. Comparison between synthetic images with and without Canny.


Figure 5. Comparison between synthetic images with and without IP-Adapter.

rear of a vehicle, indicating a conflict with the leading vehicle. A synthetic frame generated without Canny ControlNet produces a vehicle facing the wrong direction (front view), distorting the event context. Using Canny ControlNet ensures that the generated vehicle correctly maintains its rear-facing orientation, preserving the integrity of the original scene. The synthetic frame with Canny ControlNet also retains critical details, such as the status of brake lights, which accurately reflect the original frame—an essential factor for driving safety evaluation.

**IP-Adapter ControlNet**. After generating the synthetic frame at the "Impact" timestamp, the remaining five keyframes are generated while maintaining visual consistency. To achieve this, IP-Adapter is utilized to enforce structural and stylistic coherence across frames. Figure 5 presents a comparison of generated keyframes with and without IP-Adapter. In the "Impact" frame, the front vehicle is black. With IP-Adapter, subsequent frames retain this attribute, preserving key scene elements. However, without IP-Adapter, visual inconsistencies emerge—e.g., in the "Event Start" frame, the front vehicle changes to red, disrupting continuity.

## 5. Tasks and Benchmarks

The multimodal nature of SynSHRP2 supports a multitude of tasks, including detection, tracking, and prediction. In this section, we present two tasks on the SynSHRP2 dataset, including SCE attribute detection and SCE scene understanding. Additionally, several benchmarks are evaluated on these two tasks. By this means, we provide the insight of SynSHRP2 and suggest avenues for future research.

### 5.1. Task 1: SCE attribute detection

**Problem Setup**. Utilizing the SynSHRP2 kinematic dataset, we present a number of benchmarks to detect SCE attributes, including three subtasks: distinguish event severity types, incident types, and conflict types, which are crucial for the safe operation of ADS and ADAS. The distribution of SCE by event attributes can be found in the appendix. In brief, there are five event severity types (combining "Crash Severity" and "Event Type"), 15 incident types (excluding category "None," "Other," and "Unknown"), and 16 conflict types. Each SCE includes 5 seconds of triaxial acceleration and speed data.

**Data Pre-processing.** The model inputs are triaxial acceleration and speed around the occurrence of an SCE. The temporal localization of each SCE is pinpointed at the "Impact" timestamp from the SynSHRP2 database, serving as the center of the SCE. A temporal window encompassing 25 kinematic data points (representing 2.5 seconds) both preceding and succeeding the "Impact" timestamp was extracted, resulting in a 5-second interval of triaxial acceleration and speed record data.

**Model Implementation**. The dataset was randomly divided into training, testing, and validation subsets in the proportion of 7:2:1. The validation set was used to tune the hyperparameters, and the evaluation performance was based on the independent testing set. The software environment was based on Python 3.11 running on Rocky Linux 9.3. The model was trained on a high-performance GPU workstation with dual Intel Xeon Gold 6442 CPUs @ 2.60 GHz, 512 GB RAM, and one Nvidia Tesla H100 80 GB GPU.

We evaluated six benchmark models for the SCE attribute detection task, including 1-D SwinTransformer [47], CNN-GRU + XGBoost [46], CNN-LSTM [2], logistic regression [56], Adaboost [35], and random forest [52]. These models have superior performance on the original SHRP 2 NDS dataset [47].

We use the following setup for each model. The 1-D Swin Transformer model employs four Swin blocks, with the attention mechanism consistently utilizing 16 heads across all blocks. The CNN-GRU + XGBoost model consists of a convolutional layer followed by multiple GRU layers, which extract representations for XGBoost to employ classification. The CNN-LSTM model consists of a convolutional layer followed by an LSTM layer, culminating in a fully connected layer for classification. The statistical metrics used for logistic regression, Adaboost, and the random

| Method | Base Models | Accuracy | mAP | AUC | Balanced Accuracy | Macro Precsion | Macro F1 |
|---|---|---|---|---|---|---|---|
| 5-way event severity type classification | | | | | | | |
| Shi et al. [47] | 1-D SwinTransformer | 0.876 | 0.594 | 0.910 | 0.582 | **0.679** | **0.619** |
| Shi et al. [46] | CNN-GRU + XGBoost | 0.869 | 0.593 | 0.923 | 0.557 | 0.633 | 0.577 |
| Arvin et al. [2] | CNN-LSTM | **0.878** | 0.582 | 0.909 | 0.570 | 0.607 | 0.582 |
| Winlaw et al. [56] | Statistical metrics + Logistic regression | 0.865 | 0.608 | 0.921 | 0.482 | 0.616 | 0.527 |
| Osman et al. [35] | Statistical metrics + Adaboost | 0.855 | 0.497 | 0.837 | 0.523 | 0.636 | 0.552 |
| Taccari et al. [52] | Statistical metrics + Random forest | 0.871 | **0.665** | **0.926** | **0.584** | 0.649 | 0.605 |
| 15-way incident type classification | | | | | | | |
| Shi et al. [47] | 1-D SwinTransformer | 0.592 | 0.295 | 0.757 | 0.285 | 0.337 | 0.294 |
| Shi et al. [46] | CNN-GRU + XGBoost | 0.594 | 0.310 | 0.829 | 0.243 | 0.345 | 0.262 |
| Arvin et al. [2] | CNN-LSTM | **0.609** | **0.324** | **0.842** | **0.296** | 0.326 | **0.296** |
| Winlaw et al. [56] | Statistical metrics + Logistic regression | 0.589 | 0.259 | 0.827 | 0.208 | 0.244 | 0.192 |
| Osman et al. [35] | Statistical metrics + Adaboost | 0.557 | 0.156 | 0.620 | 0.186 | 0.163 | 0.168 |
| Taccari et al. [52] | Statistical metrics + Random forest | 0.607 | 0.320 | 0.829 | 0.261 | **0.416** | 0.264 |
| 16-way conflict type classification | | | | | | | |
| Shi et al. [47] | 1-D SwinTransformer | 0.581 | 0.226 | 0.774 | 0.188 | 0.197 | 0.180 |
| Shi et al. [46] | CNN-GRU + XGBoost | 0.566 | 0.247 | 0.794 | 0.206 | 0.268 | 0.212 |
| Arvin et al. [2] | CNN-LSTM | 0.585 | 0.256 | **0.826** | **0.211** | **0.289** | **0.213** |
| Winlaw et al. [56] | Statistical metrics + Logistic regression | 0.576 | 0.222 | 0.809 | 0.171 | 0.219 | 0.163 |
| Osman et al. [35] | Statistical metrics + Adaboost | 0.535 | 0.159 | 0.733 | 0.158 | 0.141 | 0.146 |
| Taccari et al. [52] | Statistical metrics + Random forest | **0.590** | **0.258** | 0.805 | 0.193 | 0.282 | 0.188 |

Table 1. Benchmark comparison in SCE attribute detection.

forest model include mean, standard deviation, maximum, minimum, and the 25th, median, and 75th percentiles of the extracted kinematic data.

All benchmark models were trained from scratch, with batch sizes optimized for one Tesla H100 GPU. The best validation accuracy epoch was selected for testing on an independent set. The optimization was conducted via Adam with an initial learning rate of 3e-4, with a cosine learning rate scheduler to refine the learning as the model converges, continuing until a minimum in validation loss is observed.

**Benchmark Comparison**. Six metrics were used to evaluate benchmark performance: accuracy, mean average precision (mAP), area under the receiver operating characteristic curve (AUC), balanced accuracy, macro precision, and macro F1. The latter three metrics focus on the imbalanced category scenarios.

Table 1 presents the results for the three subtasks. Overall, deep learning models, particularly CNN-LSTM and 1-D Swin Transformer, demonstrated strong performance across subtasks, outperforming traditional statistical methods in most metrics. For five-way event severity classification, the 1-D Swin Transformer achieved the highest macro precision (0.679) and macro F1 (0.619), while the statistical metrics + random forest led in mAP (0.665), AUC (0.926), and balanced accuracy (0.584). In 15-way incident type classification, CNN-LSTM performed best in accuracy (0.609), mAP (0.324), AUC (0.842), and balanced accuracy (0.296), whereas the random forest model attained the highest macro precision (0.416). Similarly, for 16-way conflict type classification, CNN-LSTM dominated AUC (0.826), balanced accuracy (0.211), macro precision (0.289), and macro F1 (0.213), while random forest led in accuracy (0.590) and mAP (0.258).

### 5.2. Task 2: SCE scene understanding

**Problem Setup.** Utilizing the SynSHRP2 synthetic image dataset and annotated ground truth narratives, we benchmark several VLMs to generate narrative descriptions of SCEs, which are vital for understanding SCE scenes. To ensure data quality, we continuously verify keyframes and manually annotate these ground truth narrative descriptions. The dataset version used for this task is the one available as of February 24, 2025.

**Model Implementation**. To mitigate hallucinations by VLMs, we combine SCE attribute information into the prompt for VLMs, which has proven to be an effective approach in such tasks Shi et al. [48]. Specifically, the narrative is generated using the user prompt: "Describe this driving event without personally identifiable information in one paragraph, including environment, [Event severity type], and [Conflict type]." The [Event severity type] and [Conflict type] come from the detailed annotations of SCEs.

We evaluated six state-of-the-art VLM benchmarks for the SCE scene understanding task, including Llama 3.2-Vision [16], LLaVA-Llama3 [9], MiniCPM-V [63], LLaVA [33], LLaVA-Phi3 [41], and Moondream2 [53]. To make a fair comparison, all VLM benchmarks were not fine-tuned and used with their default setup. The narratives are generated by the same prompt and evaluated quantitatively by comparing them to the event narrative of SynSHRP2.

**Benchmark Comparison**. The generated narratives are evaluated using four types of metrics, including BLEU-4

| Model | Size | BLEU-4 | ROUGE-L precision | ROUGE-L recall | ROUGE-L F1 | METEOR | BERT precision | BERT recall | BERT F1 |
|---|---|---|---|---|---|---|---|---|---|
| Llama 3.2-Vision [1] | 11B | **0.017** | 0.138 | **0.253** | 0.174 | **0.229** | 0.534 | **0.598** | 0.564 |
| LLaVA-Llama3 [9] | 8B | 0.011 | 0.118 | 0.236 | 0.156 | 0.193 | 0.521 | 0.565 | 0.542 |
| MiniCPM-V [63] | 8B | 0.011 | 0.138 | 0.202 | 0.156 | 0.209 | 0.550 | 0.596 | **0.571** |
| LLaVA [33] | 7B | 0.011 | 0.161 | 0.210 | **0.176** | 0.195 | **0.555** | 0.582 | 0.568 |
| LLaVA-Phi3 [41] | 3.8B | 0.012 | 0.147 | 0.212 | 0.167 | 0.205 | **0.555** | 0.586 | 0.570 |
| Moondream2 [53] | 1.8B | 0.011 | **0.175** | 0.151 | 0.160 | 0.147 | 0.523 | 0.506 | 0.514 |

Table 2. Benchmark comparison in SCE scene understanding.

[36], ROUGE-L [31], METEOR [3], and BERTScore [68]. To comprehensively evaluate the generative narratives relative to the ground truth, precision, recall, and F1 scores from ROUGE-L and BERTScore are used, providing a balanced measure.

As shown in Table 2, larger models like Llama 3.2-Vision and LLaVA variants performed better in recall-based metrics, while some smaller models exhibited competitive precision. Llama 3.2-Vision (11B) achieved the highest BLEU-4 score (0.017), ROUGE-L recall (0.253), METEOR (0.229), and BERT recall (0.598), demonstrating strong recall and overall SCE scene understanding performance. LLaVA (7B) led in ROUGE-L F1 (0.176) and BERT precision (0.555), while MiniCPM-V (8B) attained the highest BERT F1 score (0.571), indicating a balanced precision-recall trade-off. Moondream2 (1.8B) excelled in ROUGE-L precision (0.175) but had lower recall scores.

## 6. Conclusion

This paper introduces SynSHRP2, a publicly available synthetic multimodal driving dataset containing 1,874 crash and 6,924 near-crash events derived from the SHRP 2 NDS dataset. SynSHRP2 features de-identified keyframes generated by Stable Diffusion and ControlNet, ensuring the preservation of critical safety-related information while eliminating personally identifiable information. Additionally, SynSHRP2 includes detailed annotations on SCE attributes, environmental and traffic conditions, and time-series kinematic data spanning 5 seconds before and during each SCE. Synchronized synthetic keyframes of video and SCE narratives further enhance the usability of SynSHRP2. The method and implementation details for synthesizing the dataset are provided. Two benchmarks for SCE attribute classification and scene understanding are presented to demonstrate the potential applications of SynSHRP2 in advancing safety research and ADS development.

By publicly releasing this dataset for research, we aim to advance studies on realistic driving scenarios and traffic SCEs using NDS data, as well as support the development of safe ADS. Future work will focus on synthesizing de-identified NDS video datasets.

## References

[1] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. Accessed: 2025-02-21. 8

[2] Ramin Arvin, Asad J Khattak, and Hairong Qi. Safety critical event prediction through unified analysis of driver and vehicle volatilities: Application of deep learning methods. *Accident Analysis & Prevention*, 151:105949, 2021. 2, 6, 7

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 8

[4] Wentao Bao, 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*, pages 2682–2690, 2020. 2

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[6] Joao Carvalho, An T Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1916–1923. IEEE, 2023. 3

[7] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pages 136–153. Springer, 2017. 2

[8] comfyanonymous. Comfyui: A powerful and modular stable diffusion gui and backend, 2025. Accessed: 2025-02-22. 5

[9] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 7, 8

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2

[11] cubiq. Comfyui ipadapter plus, 2025. Accessed: 2025-02-22. 5

[12] Jean-Emmanuel Deschaud. Kitti-carla: a kitti-like dataset generated by carla simulator. *arXiv preprint arXiv:2109.00892*, 2021. 2

[13] Jean-Emmanuel Deschaud, David Duque, Jean Pierre Richa, Santiago Velasco-Forero, Beatriz Marcotegui, and François Goulette. Paris-carla-3d: A real and synthetic outdoor point cloud dataset for challenging tasks in 3d mapping. *Remote Sensing*, 13(22):4713, 2021. 2

[14] Thomas A. Dingus, Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016. 1

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[16] Fannovel16. Comfyui-controlnet-aux, 2025. Accessed: 2025-02-22. 5, 7

[17] Federal Highway Administration. SHRP2 Annual Report 2017. Technical report, 2017. 1

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 1, 2

[19] Ian Goodfellow. Deep learning, 2016. 2

[20] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. 4

[21] Yuhang Han, Zhengtao Liu, Shuo Sun, Dongen Li, Jiawei Sun, Chengran Yuan, and Marcelo H Ang Jr. Carla-loc: synthetic slam dataset with full-stack sensor setup in challenging weather and dynamic environments. *arXiv preprint arXiv:2309.08909*, 2023. 2

[22] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Technical report, Virginia Tech Transportation Institute, 2016. 1, 2

[23] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[24] Ashesh Jain, Hema S Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. *arXiv preprint arXiv:1601.00740*, 2016. 1, 2

[25] John6666. Real-archvis-xl-xlv10-sdxl, 2025. Accessed: 2025-02-22. 5

[26] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024. 3

[27] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model. In *International conference on machine learning*, pages 17920–17930. PMLR, 2023. 2

[28] Trung-Nghia Le, Shintaro Ono, Akihiro Sugimoto, and Hiroshi Kawasaki. Attention r-cnn for accident detection. In *2020 IEEE intelligent vehicles symposium (IV)*, pages 313–320. IEEE, 2020. 2

[29] Jonghyun Lee, Hansam Cho, Youngjoon Yoo, Seoung Bum Kim, and Yonghyun Jeong. Compose and conquer: diffusion-based 3d depth aware composable image synthesis. *arXiv preprint arXiv:2401.09048*, 2024. 3

[30] Haoran Li, Yaocheng Zhang, Haowei Wen, Yuanheng Zhu, and Dongbin Zhao. Stabilizing diffusion model for robotic control with dynamic programming and transition feasibility. *IEEE Transactions on Artificial Intelligence*, 2024. 3

[31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 8

[32] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pages 430–448. Springer, 2024. 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 7, 8

[34] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[35] Osama A Osman, Mustafa Hajij, Peter R Bakhit, and Sherif Ishak. Prediction of near-crashes from observed vehicle kinematics using machine learning. *Transportation Research Record*, 2673(12):463–473, 2019. 2, 6, 7

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 8

[37] Xishuai Peng, Yi Lu Murphey, Ruirui Liu, and Yuanxiang Li. Driving maneuver early detection via sequence learning from vehicle signals and video images. *Pattern Recognition*, 103:107276, 2020. 2

[38] Miguel Perez, Shane MCLAUGHLIN, Takayuki Kondo, Jonathan Antin, Julie McClafferty, Suzanne Lee, Jonathan Hankey, and Thomas Dingus. Transportation safety meets big data: the shrp 2 naturalistic driving database. *Journal of the Society of Instrument and Control Engineers*, 55(5): 415–421, 2016. 1

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[40] Veronica Radu, Mihai Nan, Mihai Trăscău, David Traian Iancu, Alexandra Ştefania Ghiţă, and Adina Magda Florea. Car crash detection in videos. In *2021 23rd International Conference on Control Systems and Computer Science (CSCS)*, pages 127–132. IEEE, 2021. 2

[41] Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S. Khan. Llava++: Extending visual capabilities with llama-3 and phi-3, 2024. 7, 8

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[43] Weibin Rong, Zhanjing Li, Wei Zhang, and Lining Sun. An improved canny edge detection algorithm. In *2014 IEEE international conference on mechatronics and automation*, pages 577–582. IEEE, 2014. 4

[44] John M Scanlon, Kristofer D Kusano, Laura A Fraade-Blanar, Timothy L McMurry, Yin-Hsiu Chen, and Trent Victor. Benchmarks for retrospective automated driving system crash rate analysis using police-reported crash data. *Traffic Injury Prevention*, 25(sup1):S51–S65, 2024. 1

[45] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 2

[46] Liang Shi, Chen Qian, and Feng Guo. Real-time driving risk assessment using deep learning with xgboost. *Accident Analysis & Prevention*, 178:106836, 2022. 2, 6, 7

[47] Liang Shi, Yixin Chen, Meimei Liu, and Feng Guo. Dust: Dual swin transformer for multi-modal video and time-series modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 6, 7

[48] Liang Shi, Boyu Jiang, Tong Zeng, and Feng Guo. Scvlm: Enhancing vision-language model for safety-critical event understanding, 2025. 7

[49] Matteo Simoncini, Douglas Coimbra de Andrade, Leonardo Taccari, Samuele Salti, Luca Kubin, Fabio Schoen, and Francesco Sambo. Unsafe maneuver classification from dashcam video and gps/imu sensors using spatio-temporal attention selector. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15605–15615, 2022. 2

[50] Pierre-Louis Wolfgang Léon Suckrow, Christoph Johannes Weber, and Sylvia Rothe. Diffusion-based sound synthesis in music production. In *Proceedings of the 12th ACM SIGPLAN International Workshop on Functional Art, Music, Modelling, and Design*, pages 55–64, 2024. 2

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[52] Leonardo Taccari, Francesco Sambo, Luca Bravi, Samuele Salti, Leonardo Sarti, Matteo Simoncini, and Alessandro Lori. Classification of crash and near-crash events from dashcam videos and telematics. In *2018 21st International Conference on intelligent transportation systems (ITSC)*, pages 2460–2465. IEEE, 2018. 2, 6, 7

[53] Vikhyat. Moondream2, 2024. Accessed: 2025-02-21. 7, 8

[54] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 4, 5

[55] Joey Wilson, Jingyu Song, Yuewei Fu, Arthur Zhang, Andrew Capodieci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. Motionsc: Data set and network for real-time semantic mapping in dynamic environments. *IEEE Robotics and Automation Letters*, 7(3):8439–8446, 2022. 2

[56] Manda Winlaw, Stefan H Steiner, R Jock MacKay, and Allaa R Hilal. Using telematics data to find risky driver behaviour. *Accident Analysis & Prevention*, 131:131–136, 2019. 2, 6, 7

[57] WSJUSA. Comfyui-stablesr: Integrating stablesr into comfyui, 2025. Accessed: 2025-02-22. 5

[58] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024. 2

[59] Longfei Xu, Hongbo Huang, Yushuang Zhao, Shuwen Pan, Yaolin Zheng, Xiaoxu Yan, Linkai Huang, and Lishan Wu. Fine-grained image editing using controlnet: Expanding possibilities in visual manipulation. In *International Conference on Intelligent Computing*, pages 27–38. Springer, 2024. 3

[60] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2

[61] Hyemin Yang, Heekyung Yang, and Kyungha Min. Artfusion: A diffusion model-based style synthesis framework for portraits. *Electronics*, 13(3):509, 2024. 3

[62] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280. IEEE, 2019. 2

[63] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. 7, 8

[64] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 4, 5

[65] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 2

[66] Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In

*Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7226–7236, 2023. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4

[68] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 8

[69] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 2

[70] Jannik Zürn, Paul Gladkov, Sofía Dudas, Fergal Cotter, Sofi Toteva, Jamie Shotton, Vasiliki Simaiaki, and Nikhil Mohan. Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving. *arXiv preprint arXiv:2407.08280*, 2024. 2

# Appendix: Variable dictionary

Table 3. Variable dictionary for sensor data.

| Variable | Definition | Unit/Category | Frequency |
|---|---|---|---|
| Longitudinal acceleration | Vehicle acceleration in the longitudinal direction versus time. | g | 10Hz |
| Lateral acceleration | Vehicle acceleration in the lateral direction versus time. | g | 10Hz |
| Vertical acceleration | Vehicle acceleration vertically (up or down) versus time. | g | 10Hz |
| Speed | Vehicle speed indicated on speedometer collected from network. | km/h | 10Hz |
| Pedal brake state | On or off press of brake pedal. | 0=off, 1=on, 2=invalid data, 3=data not available | Varies |
| Lane width | Distance between the inside edge of the innermost lane marking to the left and right of the vehicle. | cm | 30Hz |
| Left line right distance | Distance from vehicle centerline to inside of left side lane marker based on vehicle based machine vision. | cm | 30Hz |
| Right line left distance | Distance from vehicle centerline to inside of right side lane marker based on vehicle based machine vision. | cm | 30Hz |

Table 4. Variable dictionary for tabular records

| Variable | Definition | Category | Count |
|---|---|---|---|
| Event Type | The outcome of each event. | Crash | 1874 |
| | | Near-Crash | 6924 |
| Crash severity | A ranking of crash severity based on vehicle dynamics, property damage, injury data, and risk to road users. | IV - Low-risk Tire Strike | 800 |
| | | III - Minor Crash | 777 |
| | | II - Police-reportable Crash | 183 |
| | | I - Most Severe | 114 |
| Incident type | The subject vehicle's conflict type in the most severe incident. | Rear-end, striking | 3916 |
| | | Road departure (left or right) | 1262 |
| | | Sideswipe, same direction (left or right) | 1052 |
| | | Turn into path (same direction) | 397 |
| | | Animal-related | 372 |
| | | Turn into path (opposite direction) | 328 |
| | | Turn across path | 328 |
| | | Rear-end, struck | 205 |
| | | Straight crossing path | 178 |
| | | Pedestrian-related | 169 |
| | | Road departure (end) | 137 |
| | | Backing into traffic | 119 |
| | | Opposite direction (head-on or sideswipe) | 91 |
| | | Backing, fixed object | 87 |
| | | Pedalcyclist-related | 67 |
| Conflict type | The note about the other object(s) involved in the incident. | Conflict with a lead vehicle | 3290 |
| | | Conflict with vehicle in adjacent lane | 1571 |
| | | Single vehicle conflict | 1479 |
| | | Conflict with vehicle turning into another vehicle path (same direction) | 394 |
| | | Conflict with animal | 372 |
| | | Conflict with vehicle turning into another vehicle path (opposite direction) | 326 |
| | | Conflict with vehicle turning across another vehicle path (opposite direction) | 253 |
| | | Conflict with obstacle/object in roadway | 187 |
| | | Conflict with a following vehicle | 186 |
| | | Conflict with parked vehicle | 175 |
| | | Conflict with vehicle moving across another vehicle path (through intersection) | 175 |
| | | Conflict with pedestrian | 169 |
| | | Conflict with merging vehicle | 131 |
| | | Conflict with oncoming traffic | 90 |
| | | Conflict with vehicle turning across another vehicle path (same direction) | 74 |
| | | Conflict with pedal cyclist | 67 |