Automated Construction of Living Mobility Datasets from Public Video: A YouTube Case Study

MD SHADAB ALAM*, Eindhoven University of Technology, The Netherlands
MARIUS HOGGENMUELLER*, University of Sydney, Australia
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands

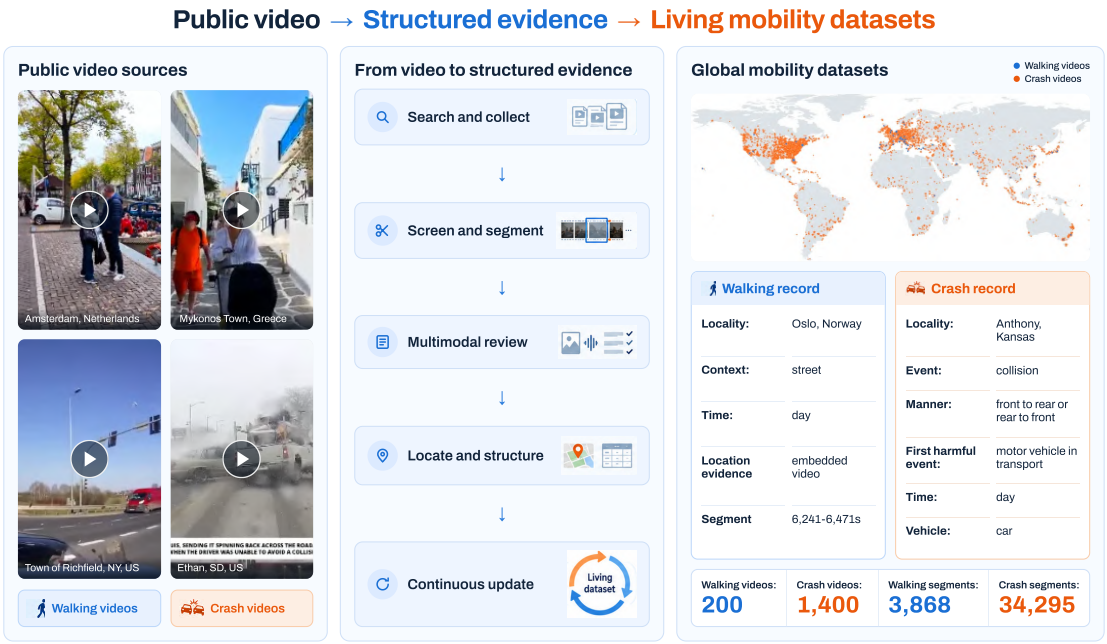

Fig. 1. **Overview of our living mobility dataset infrastructure.** Publicly shared walking and crash videos are transformed into structured research data through automated collection, screening, temporal segmentation, multimodal review, geographic grounding, and continuous updating. The two use cases demonstrate how the same infrastructure supports both everyday pedestrian environments and road traffic crashes and near collisions. The source video examples are illustrative and do not indicate restrictions on video orientation, format, or duration, while the map shows the geographic distribution of resolved locality reference coordinates in the current corpora.

Public video contains naturalistic mobility footage, but transforming it into reusable research data requires more than a search and download. We present an automated and resumable framework for constructing living mobility datasets from YouTube, instantiated through two domain specific pipelines: pedestrian walking environments and road traffic crashes and near collisions. The pipelines combine discovery, Qwen metadata screening, GPU temporal segmentation, Cosmos3 Nano visual review, geographic grounding, provenance tracking, and versioned output. The resulting corpora contain 3,868 walking segments and 34,295 crash related segments. The walking corpus includes 89 resolved locality records in 28 countries, while the crash corpus contains 17,588 geographically resolved segments that represent 5,270 resolved locality records. Because the framework is resumable and continuously executable,

Authors' Contact Information: Md Shadab Alam, m.s.alam@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands; Marius Hoggenmueller, marius.hoggenmueller@sydney.edu.au, University of Sydney, Sydney, Australia; Pavlo Bazilinskyy, p.bazilinskyy@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands.

the collection can run to expand datasets over time while preserving traceable processing decisions and reproducible snapshots. Its modular design can be adapted to other large scale video collections beyond mobility and YouTube.

Additional Key Words and Phrases: mobility datasets, public video platforms, YouTube, traffic safety, vision language models, dataset infrastructure

1 Introduction

Road traffic and daily mobility unfold in complex social and physical environments. Approximately 1.19 million people die in traffic crashes every year worldwide, with pedestrians, cyclists, and motorcyclists accounting for more than half of global road fatalities [30]. Beyond crashes and near collisions, people constantly navigate streets, intersections, public squares, markets, parks, and other shared spaces shaped by surrounding road users, infrastructure, environmental conditions, and local context. Understanding these situations is relevant for transportation, computer vision (CV), and human computer interaction (HCI), as intelligent mobility systems increasingly mediate how people perceive, navigate, and respond to their surroundings.

Large visual datasets have become central to computational mobility research. KITTI and Cityscapes established influential benchmarks for automated driving and understanding of urban scenes [9, 13], while BDD100K, nuScenes, and the Waymo Open Dataset expanded environmental, geographic, and sensor diversity [6, 29, 32]. Pedestrian focused resources such as PIE support modelling of pedestrian actions and trajectories [27]. Safety focused datasets complement these resources: DADA 2000 combines accident footage with driver attention information [11], the Car Crash Dataset provides real world collision videos for accident anticipation [4], and the Nexar Dashcam Collision Prediction Dataset includes collisions, near collisions, and ordinary driving [20]. A global dataset of continuous urban dashcam driving shows how publicly available video can provide geographically diverse observations of everyday road environments [2]. Together, these resources demonstrate the value of structured real world mobility observations, but individual datasets remain bounded by their recording systems, locations, viewpoints, selection criteria, and collection procedures.

YouTube provides a complementary source of naturalistic observations because users routinely share videos documenting everyday activities, environments, journeys, and experiences. These include walking tours, driving and dashcam footage, cycling videos, and recordings of public spaces [3]. Mobility research has already demonstrated the value of such material as an observational source. Chambers et al. used YouTube segments to study walking synchronisation under naturalistic conditions [8], while Hartmann and Purves showed that city walking tour videos provide rich street level information on pedestrian and cyclist movement patterns [14]. Beyond mobility, digital ethnography has used user generated YouTube videos to investigate naturally occurring interactions between people and service robots in public places [22], while related online ethnography has examined encounters between delivery robots and road and sidewalk users through user generated social media videos [33]. Together, these studies show that online video can provide access to naturally occurring behaviours, interactions, and environments that may be difficult to capture systematically through laboratory or researcher controlled data collection.

Yet this abundance of publicly available video also creates a methodological challenge: publicly available online video is not, by itself, a research dataset, and the scale of contemporary platforms makes comprehensive manual curation increasingly impractical. Large scale video resources illustrate the magnitude of this problem: the large-scale labeled video dataset YouTube 8M was constructed from approximately eight million YouTube videos representing around 500,000 hours of footage [1]; similarly HowTo100M collected 136 million clips from 1.22 million narrated web videos [19]. In subsequent work, it has been noted that manual video annotation is expensive and time consuming on

such scales, motivating increasingly automated approaches to data collection and organisation [19]. Public video also presents additional challenges for research use: search results can be noisy, uploads vary in relevance and duration, useful observations may occupy only part of a video, and metadata may be incomplete or inconsistent. Videos may also be duplicated across searches or become unavailable over time. Therefore, transforming this material into reusable research data requires scalable discovery, screening, temporal segmentation, semantic description, geographic grounding, deduplication, provenance tracking, and explicit treatment of uncertain evidence.

Recent advances in vision language models (VLMs) make it increasingly practical to automate video understanding at a scale that would be difficult to achieve through manual curation alone. Multimodal models can interpret temporal, semantic, and contextual information beyond conventional object detection. Driving orientated systems such as DriveLM organise multimodal reasoning around perception, prediction, and planning [28], while LingoQA demonstrates video question answering for complex driving situations [18]. These capabilities enable a shift from manually assembled video collections towards automated pipelines that can discover, screen, segment, and structure newly available material at scale. In our setting, model generated outputs are retained together with their provenance and processing history, allowing the resulting dataset records to remain traceable across repeated and continuously expanding collection runs.

These requirements connect directly to the broader principles of dataset documentation and responsible use. Data Statements [5], Datasheets for Datasets [12], and Data Cards [25] emphasise transparent reporting of dataset motivation, composition, provenance, collection procedures, intended use, and limitations. For automatically constructed datasets, traceability additionally requires maintaining the connection between each resulting record and the processing decisions that produced it. In our setting, this includes the source video, screening decisions, temporal boundaries, model generated semantic descriptions, geographic evidence and resolution outcomes, and the versioned dataset snapshot in which the record appears. Preserving this information allows researchers to inspect how individual records were derived rather than treating structured annotations as detached outputs.

A further challenge is that YouTube changes continuously, whereas research datasets are commonly treated as fixed artefacts. New videos appear, existing videos become unavailable, and repeated discovery can expose material that was absent from earlier searches. Dynamic resources such as the continuously growing corpus of Lan et al. [17], Dynabench [15], and CORE [16] demonstrate that datasets can instead function as maintained research infrastructure. For public video, this creates an important tension between growth and reproducibility. A collection pipeline should be able to continue running so that new observations can be incorporated over time, while immutable, versioned snapshots are required to preserve the exact state used for a published analysis.

Together, these developments motivate YouTube not only to treat it as a repository from which individual clips are manually selected but as a continuously changing source of naturalistic observations that can support systematic research at scale. The methodological challenge to do so is to convert large and heterogeneous collections of user generated video into structured research resources while preserving provenance, reproducibility, and the ability to incorporate newly available material over time. While the present work addresses this challenge in the context of mobility research, we foresee the approach being useful across a much wider range of domains, including online video-based observational and ethnographic research.

1.1 Aim of the study

The aim of this study is to develop and demonstrate an automated, scalable, and resumable framework to construct live research datasets from YouTube. We instantiate the framework through two complementary mobility use cases: everyday pedestrian environments and crash related road traffic events, including crashes and near collisions. The

corresponding pipelines combine automated discovery, metadata screening, temporal segmentation, multimodal review, semantic annotation, geographic grounding, provenance tracking, and versioned outputs. Repeated execution allows the resulting corpora to expand as new material becomes available, while preserving reproducible snapshots and the processing history of individual records. Although demonstrated through mobility, the framework is designed to be adaptable to other research settings that rely on large collections of publicly available video. Figure 1 provides an overview of how the two mobility use cases are organised within the shared infrastructure.

2 Method

2.1 Framework Overview

Although the framework can be applied to other online video collections, the present study implements and evaluates it using YouTube. Figure 2 shows the processing stages shared by the two use cases. The framework is designed around a simple research need: public video is abundant, but individual uploads cannot be treated directly as research observations. Search results may be irrelevant, long uploads may contain several recordings or unrelated material, useful footage may occupy only part of a video, and previously available sources may later disappear. We therefore organise collection as a sequence of stages that progressively turn public videos into traceable research records: discovery, metadata screening, media acquisition, temporal decomposition, visual review, geographic grounding, and structured dataset generation.

The same overall structure is used for both corpora, but the decisions made within it reflect the research question. The walking pipeline seeks sustained pedestrian movement through real environments, whereas the crash and near collision pipeline seeks visible road traffic safety events. At each stage, the framework records what was decided and which evidence supported that decision. This allows an interrupted collection to continue from the point it reached, prevents completed work from being unnecessarily repeated, and makes it possible to revisit one part of the process when an annotation procedure changes. The resulting dataset is therefore not only a collection of accepted clips, but also a record of how those clips were selected, divided, described, and geographically grounded.

Different components are used according to the type of evidence required. Qwen3-4B-Instruct-2507 screens uploader supplied text before video analysis [26]. FFmpeg supports media processing and proposes visual transitions that may indicate boundaries between recordings. NVIDIA Cosmos3-Nano then reviews visual and temporal evidence for decisions that require inspection of the video itself [24]. Candidate transitions are not accepted automatically as final boundaries; they are checked during multimodal review because camera motion, exposure changes, and occlusion can resemble edits. This division of responsibilities reduces unnecessary video analysis while keeping visually dependent decisions tied to the source footage.

2.2 Candidate Discovery, Metadata Screening, and Media Acquisition

Candidate discovery and metadata screening correspond to Steps 1 and 2 in Figure 2. The same general procedure is used for both corpora, while the search terms, duration criteria, and screening thresholds are configured separately and summarised in Table 1. Videos are discovered through the YouTube Data API using predefined queries. Collection proceeds in manageable batches, and the framework records both the search position and videos that have already been encountered. This allows discovery to resume after interruption without repeatedly introducing the same source videos. The duration criteria reflect the source material expected in each use case: walking videos are generally long

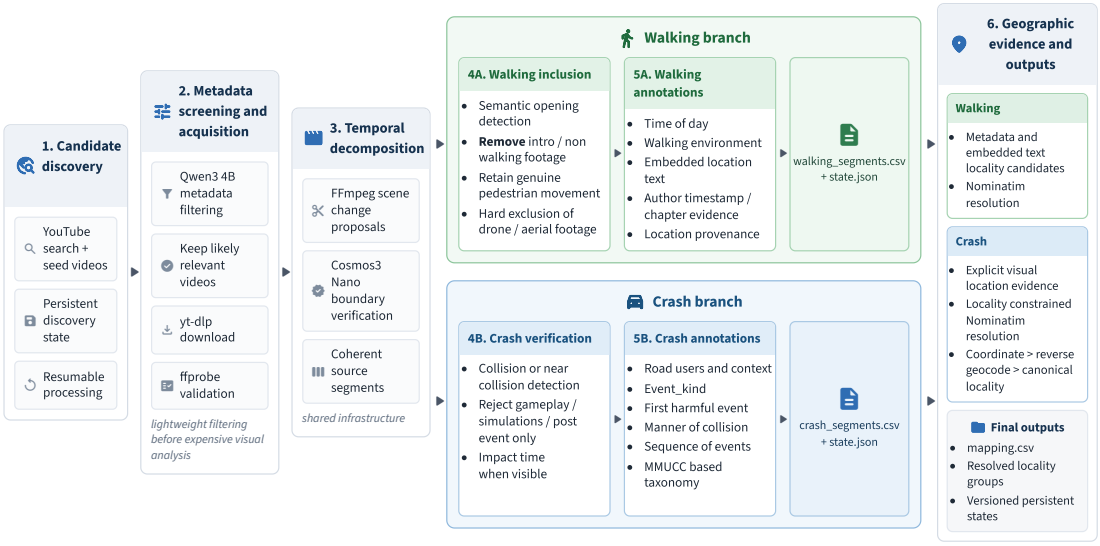

Fig. 2. Overview of the continuous YouTube mobility dataset framework for the walking and crash video pipelines.

observations of movement through an environment, whereas crash and near collision footage may occur either in short single event uploads or in longer compilations.

Before a video is downloaded, Qwen3-4B-Instruct-2507 evaluates the uploader supplied title and description [26]. This stage is intentionally an early plausibility check rather than a final inclusion decision. Its purpose is to avoid spending substantially more computation on videos whose own metadata provides little indication that they contain the target material. Candidates that pass this screen still require visual review. The walking and crash pipelines use different operational thresholds because their search results contain different types of irrelevant material; the values and prompt exclusions are reported in Table 1. These confidence values are used as screening criteria rather than as calibrated probabilities.

Videos that pass metadata screening are downloaded temporarily using `yt-dlp` (https://github.com/yt-dlp/yt-dlp) and checked with `ffprobe` (https://ffmpeg.org/ffprobe.html) before visual processing. A valid local copy is reused when available. The source media are retained only while needed for processing, while the decisions, annotations, provenance, and resulting dataset records are preserved independently of the downloaded video file.

2.3 Temporal Decomposition and Visual Inclusion

Both pipelines separate temporal decomposition from semantic inclusion. Potential edit boundaries are first proposed using FFmpeg based scene change detection. These proposed cuts are not accepted directly because rapid turns, camera shake, exposure changes, occlusion, and other continuous motion can resemble an edit. Instead, Cosmos automatically reviews a short clip around each proposed boundary and determines whether the adjacent visual evidence supports a genuine discontinuity between scenes or source recordings. Subsequently, only verified edit boundaries are used to define segments. The use case specific parameters and decision thresholds for this process are reported separately in the corresponding pipeline configurations.

Table 1. Use case specific configuration of candidate discovery and metadata screening.

Configuration	Walking environments	Road traffic crashes and near collisions
Search queries	“walking tour”; “city walk”; “walk with me”; “virtual walk”; “4k walking tour”; “POV walking”; “street walk”; “nature walk”; “rain walk”; “night walk”	“car crash compilation dashcam”; “traffic accident compilation dashcam”; “road crash compilation CCTV”; “car accidents caught on camera”
Maximum search pages per query and discovery cycle	2	4
Results requested per API page	50	50
Maximum new candidates per discovery batch	200	200
Video duration criterion	Minimum 300 s	No minimum or maximum duration restriction
Metadata input	Video title and first 4,000 characters of the description	Video title and first 5,000 characters of the description
Screening target	Likely pedestrian walking through real physical environments	Likely genuine road traffic crashes or near collisions
Prompt exclusions	Workouts, treadmills, games, animation, music videos, product reviews, driving or cycling tours, public transport rides, drone footage, flights, boat rides, talking head videos, slideshows, maps, static ambience, and unrelated uses of “walking”	Video games, simulations, films, television drama, animation, crash tests, motorsport only footage, toy vehicles, commentary without original footage, emergency response footage without the event, and unrelated uses of “crash”
Geographic fields extracted	Locality, state, and country when directly supported by metadata	Locality, state, and country when directly supported by metadata
Metadata confidence threshold	0.60	0.65
Additional source metadata enrichment	Title, upload date, channel identifier, view count, description, chapters, represented segment count, and metadata update date	Title, upload date, channel identifier, view count, description, chapters, represented segment count, and metadata update date

Walking video inclusion. The walking pipeline includes an additional automated stage to identify where sustained walking content begins. Cosmos reviews the opening portion of each accepted upload and estimates the start of the main walking content, allowing introductory material, highlights, and other preceding footage to be excluded before segment level analysis. The resulting semantic opening boundary is combined with the verified edit boundaries to define candidate walking intervals. Longer intervals are divided into bounded review windows, while intervals that are too short for reliable assessment are excluded.

Cosmos automatically evaluates each candidate walking segment using short continuous motion samples drawn from the beginning, middle, and end of the source interval. The model determines whether the footage represents real walking through a real physical environment by assessing whether the viewpoint and motion are consistent with the height and progression of the pedestrian on foot. The automated review excludes drone and aerial footage, vehicle only movement, cycling, boating, train or bus footage, static scenes, talking head content, advertisements, channel promotion, opening highlights, title cards, games, animation, slideshows, and other non walking material. Segments that satisfy the walking inclusion criteria are retained for subsequent annotation and geographic processing.

Crash video inclusion. The crash pipeline uses the verified edit boundaries to partition each accepted upload into complete source segments. Then each segment is automatically processed. The pipeline creates a temporally sampled

representation spanning the complete source segment and provides it to Cosmos together with structured instructions for determining whether the footage contains a real road traffic collision or a near collision in which evasive action narrowly prevents impact. Cosmos is explicitly instructed to reject gameplay, simulations, staged scenes, crash tests, motorsport incidents, title cards, commentary only footage, emergency response footage without the event itself, and footage showing only post-collision damage after an unseen collision. The model returns a structured decision that indicates whether a relevant event is present, together with its confidence and associated event attributes. The pipeline automatically checks this response for syntactic and semantic consistency and retries invalid or contradictory responses. Segments are retained only when the automated Cosmos decision satisfies the inclusion criterion; otherwise, they are excluded without requiring a manual inclusion decision. For collision events, Cosmos also estimates the visible impact time relative to the source segment when this can be established from the footage.

2.4 Domain Specific Semantic Annotation

The two corpora share the same annotation principle: an attribute is recorded only when it is supported by the source material, and automatically generated descriptions are retained as structured research metadata rather than treated as manually verified ground truth. The specific annotations differ because everyday walking footage and safety critical traffic events support different research questions.

Walking corpus. Each accepted walking segment records its temporal boundaries, time of day, walking environment, location provenance, and visual confidence. Time of day is represented as day, night, dawn or dusk, or unknown. Walking environment is classified as street, indoor, beach, park or nature, trail, market, waterfront, square or plaza, transport hub, mixed, other, or unknown. Here, *other* denotes an identifiable environment that falls outside the predefined categories, whereas *unknown* is used when the environment cannot be determined reliably. The model also extracts readable location text that appears directly in the reviewed footage while excluding channel branding, generic titles, and inferred places.

Uploader supplied timestamps or chapter descriptions are aligned with the accepted segment boundaries. The dataset therefore preserves whether geographic evidence comes from an uploader supplied timestamp or chapter, readable text embedded in the video, both sources, or neither. Keeping these sources separate allows later analyses to distinguish information supplied by the uploader from information visible in the recording itself.

Crash corpus. For accepted traffic safety segments, Cosmos describes the camera perspective, visible road users, road environment, time of day, weather, road surface condition, visible outcomes, and the event itself. The annotation instructions explicitly prohibit conclusions about fault, legal responsibility, intention, impairment, distraction, injury severity, fatalities, or other characteristics that cannot be established from the footage.

Crash events are described using selected concepts from the sixth edition of the National Highway Traffic Safety Administration Model Minimum Uniform Crash Criteria (MMUCC) [21]. The annotation first distinguishes a collision from a near collision. For collisions, it records the first visually supported harmful event, the manner of collision when another motor vehicle is involved, and up to four visually supported events in chronological order. This allows a sequence such as leaving the roadway and then overturning to remain represented as two events rather than being collapsed into one broad crash category. These annotations use MMUCC concepts to provide a consistent descriptive vocabulary, but they are based on public video and are not presented as official police coded crash reports.

2.5 Geographic Evidence and Resolution

Geographic grounding is deliberately conservative in both pipelines. Inferring a place from general visual appearance is difficult across diverse environments and viewpoints [7, 31]. We therefore publish a location only when explicit evidence supports it and preserve the distinction between the original evidence and the final resolved locality.

For the walking corpus, Qwen extracts locality, state or province, and country only when these are directly supported by the video title or description. Uploader supplied timestamps or chapter labels and readable location text within the footage provide additional segment level evidence. Place names are resolved with OpenStreetMap Nominatim [23] to obtain a consistent locality representation and reference coordinates. The dataset retains where the location evidence came from, allowing researchers to distinguish uploader supplied information from text observed directly in the video.

Crash videos require a stricter approach because compilations often combine recordings from different places and may provide little useful location information in their metadata. Cosmos therefore searches accepted segments for explicit visual evidence such as location overlays, captions, road signs, business signs, and visible coordinates. The pipeline does not infer a location from scenery, architecture, language, number plates, vegetation, or similar indirect cues, and visual geographic evidence is accepted only at a minimum confidence of 0.90.

Crash place candidates are resolved with Nominatim only when they can be associated with a locality such as a city, town, village, neighbourhood, or comparable inhabited place. Roads, stations, businesses, counties, states, and unrelated points of interest are not used as substitutes for a locality. When latitude and longitude are visible in the footage, the coordinates are used to identify the containing locality. The original coordinates remain part of the evidence record, while the coordinates published in the dataset refer to the resolved locality. They should therefore not be interpreted as the exact position of the camera or event.

2.6 Maintaining an Evolving and Reproducible Dataset

A central design problem is that public video collections change while published research needs a stable and reproducible dataset. New videos can appear, existing videos can disappear, and annotation procedures can improve. The framework therefore records how far each video and segment has progressed through the collection process, together with the decisions and evidence produced along the way. If processing is interrupted, completed work remains available when collection resumes rather than requiring the corpus to be rebuilt from the beginning.

This record also allows methodological changes to be applied selectively. For example, an improved geographic procedure can be applied again while retaining previously accepted segment boundaries, and a revised crash taxonomy can be added only to segments that still require the new annotations. If an older source video is no longer available, its previously accepted record is retained and the missing new annotation is recorded as unavailable rather than inferred from earlier metadata. In this way, changes to one part of the methodology do not require unrelated parts of the dataset to be reconstructed.

The same records are used to generate the research outputs for each corpus. Walking outputs contain accepted segment boundaries, environmental and temporal annotations, geographic evidence, and source information. Crash outputs contain accepted safety events, contextual and MMUCC based annotations, and a separate geographic view for successfully resolved localities. Unresolved records remain part of the corpus rather than being assigned an artificial "unknown" place. Fixed releases can therefore capture the exact dataset used for a study, while the maintained collection continues to incorporate new material and improved annotations.

Table 2. Summary of the walking corpus snapshot.

Measure	Walking corpus
Discovered videos	200
Videos passing metadata screening	156
Metadata rejected videos	44
Completed video records	147
Visual rejected videos	8
Unique source videos represented	147
Accepted segments	3,868
Retained duration (hours)	201.29

Across both corpora, the automatically derived annotations are intended to support large scale exploration, corpus characterisation, retrieval, model development, and subsequent dataset construction. They should not be interpreted as manually verified ground truth. The MMUCC based crash variables describe visually observable characteristics rather than official MMUCC coded police reports, while published geographic coordinates represent geocoded reference localities rather than exact event positions.

3 Results

All counts reported in this section refer to the versioned dataset snapshot used for this manuscript.

3.1 Walking Corpus

The walking corpus contains 3,868 accepted segments drawn from 147 unique source videos, corresponding to 201.29 hours of retained footage. In total, 200 video records were discovered, of which 156 passed metadata screening. Of these, 147 completed the full processing workflow, 8 were rejected during visual review, and 1 video could not be downloaded. No visual, text, or pipeline errors were recorded in the analysed snapshot. Table 2 summarises the resulting walking corpus.

The accepted segments represent a range of walking environments and recording conditions. Figure 3 shows the geographic distribution of the four most prevalent walking environment categories. Street environments account for 2,272 segments (58.74%), followed by mixed environments with 595 (15.38%), indoor environments with 311 (8.04%), markets with 240 (6.20%), squares or plazas with 169 (4.37%), parks or nature with 130 (3.36%), waterfronts with 117 (3.02%), trails with 15 (0.39%), beaches with 10 (0.26%), and transport hubs with 9 (0.23%). No accepted segments were labelled as other or unknown environments in this snapshot. Daytime footage accounts for 2,787 segments (72.05%), while 662 segments (17.11%) were recorded at night and 383 (9.90%) at dawn or dusk. The time of day could not be determined for 36 segments (0.93%).

Geographic evidence was available for 3,865 of the 3,868 accepted segments (99.92%). Segment specific evidence, consisting of uploader supplied timestamp or chapter annotations, readable embedded location text, or both, was available for 2,666 segments (68.92%), while video metadata provided geographic evidence for 3,865 segments (99.92%). Of all accepted segments, 982 (25.39%) contained location evidence only from uploader supplied timestamps or chapter descriptions, 1,228 (31.75%) contained only readable location text embedded in the video, and 456 (11.79%) contained both forms of segment specific evidence. The remaining 1,202 segments (31.08%) contained neither form of segment

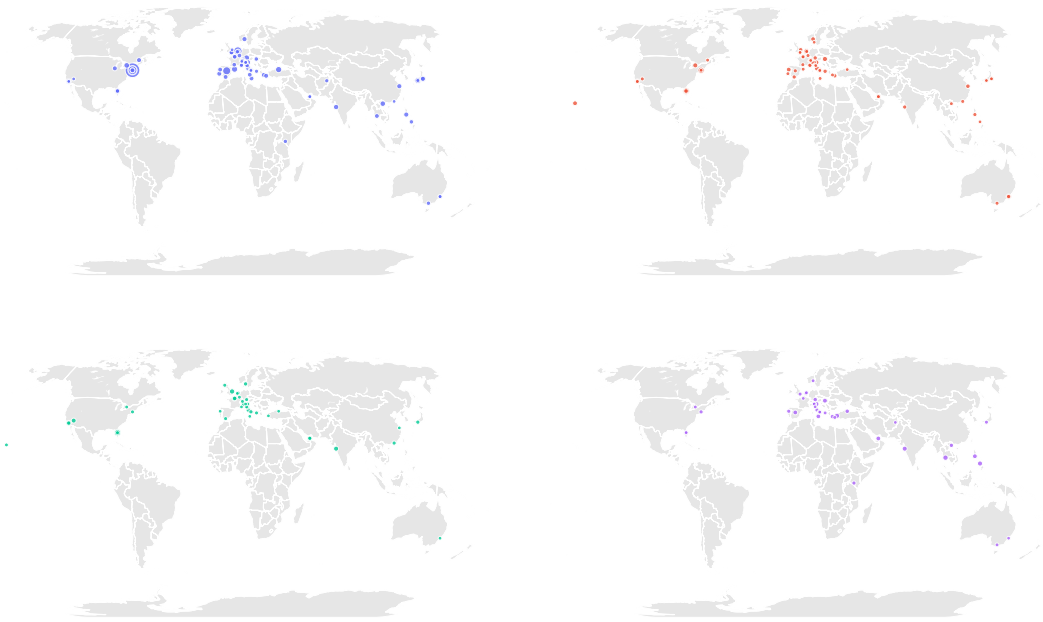

Fig. 3. Geographic distribution of walking segments by walking environment. The four panels show the most prevalent environment categories: top left, street; top right, mixed; bottom left, indoor; and bottom right, market. Across all 3,868 accepted walking segments, street environments occur in 2,272 segments (58.74%), mixed environments in 595 (15.38%), indoor environments in 311 (8.04%), and markets in 240 (6.20%).

specific evidence. Overall, timestamp or chapter evidence was available for 1,438 segments (37.18%), while embedded video location text was available for 1,684 (43.54%).

Nominatim resolution produced 3,490 segments with a resolved locality, corresponding to 90.23% of the walking corpus. The remaining 378 segments (9.77%) did not have a resolved locality: 375 returned no geocoding result and 3 did not undergo geocoding. The resolved segments represent 89 canonical localities across 28 countries or territories and five continents, originate from 129 source videos, and correspond to 180.91 hours of retained footage. North America accounts for 1,529 resolved segments (43.81%), followed by Europe with 1,387 (39.74%), Asia with 488 (13.98%), Oceania with 76 (2.18%), and Africa with 10 (0.29%). The United States is the most represented country with 1,437 resolved segments (41.17%), followed by Italy with 408 (11.69%), Spain with 216 (6.19%), the Netherlands with 193 (5.53%), and France with 101 (2.89%). The retained walking segments have a mean duration of 187.34 seconds and a median duration of 227.00 seconds, with 25th and 75th percentiles of 183.00 and 233.00 seconds, respectively, and a maximum duration of 239.00 seconds.

3.2 Crash and Near Collision Corpus

The crash corpus contains 34,295 accepted traffic safety segments corresponding to 151.35 hours of retained footage. The analysed state contains 1,400 video records, of which 920 passed metadata screening. Of the complete set of video

Table 3. Summary of the crash and near collision corpus snapshot.

Measure	Crash corpus
Video records	1,400
Videos passing metadata screening	920
Metadata rejected videos	480
Completed video records	902
Visual rejected videos	16
Visual processing errors	2
Accepted segments	34,295
Retained duration (hours)	151.35

records, 902 reached the completed processing state, 480 were rejected during metadata screening, 16 were rejected during visual review, and 2 ended with a visual processing error. Table 3 summarises the resulting crash and near collision corpus.

The structured annotations describe both the road users represented in the corpus and the recording conditions. Cars occur in 32,213 segments (93.93%), trucks in 6,596 (19.23%), vans or pickup trucks in 3,369 (9.82%), and motorcycles in 1,186 (3.46%). Pedestrians occur in 834 segments (2.43%), SUVs in 766 (2.23%), light trucks in 522 (1.52%), buses in 362 (1.06%), and cyclists in 227 (0.66%). Because several road user categories may occur within the same segment, these percentages are not mutually exclusive. Figure 4 shows the geographic distribution of the four most frequently represented road user categories. Daytime footage accounted for 27,776 segments (80.99%), night footage for 4,761 (13.88%), dawn or dusk for 1,566 (4.57%), dusk for 49 (0.14%), and 143 segments (0.42%) had an unknown time of day.

At the current snapshot, 2,884 segments have been classified using the MMUCC based taxonomy, while 31,405 remain pending and 6 ended with a terminal model error. Among the classified segments, 2,780 (96.39%) were identified as collisions and 104 (3.61%) as near collisions. Among the collision segments currently classified, 2,768 (99.57%) were classified as front to rear or rear to front collisions, 7 (0.25%) as front to front collisions and 4 (0.14%) as sideswipes in the same direction. A segment (0.04%) involved a collision that did not involve another motor vehicle in transport. The first harmful event was a motor vehicle in transport for 2,779 collision segments (99.96%) and ground for one segment (0.04%). All 2,780 currently classified collisions contained one event in the MMUCC sequence representation. Figure 5 visualises the geographic distribution of the most frequent manner of collision categories.

Geographic resolution produced 17,588 segments with a publishable locality, corresponding to 51.28% of the accepted crash corpus, while the remaining 16,707 segments (48.72%) were unresolved. The resolved segments represent 5,270 canonical locality entities distributed across six continents. North America accounts for 11,271 resolved segments (64.08%), followed by Europe with 3,441 (19.56%), Asia with 1,133 (6.44%), Oceania with 827 (4.70%), South America with 484 (2.75%), and Africa with 398 (2.26%). The United States is the country with 10,293 resolved segments (58.52%), followed by Canada with 858 (4.88%), Australia with 792 (4.50%), Italy with 504 (2.87%), and the United Kingdom with 464 (2.64%).

Among the unresolved crash segments, 9,254 had no resolvable locality result and 7,451 contained no geographic evidence, while one segment produced a rejected resolution result and one ended in a resolution failure. The 17,588 geographically resolved segments correspond to 82.59 hours of footage and originate from 714 unique source uploads. The geographic mapping contains 5,270 locality rows, and its expanded segment count matches the number of resolved

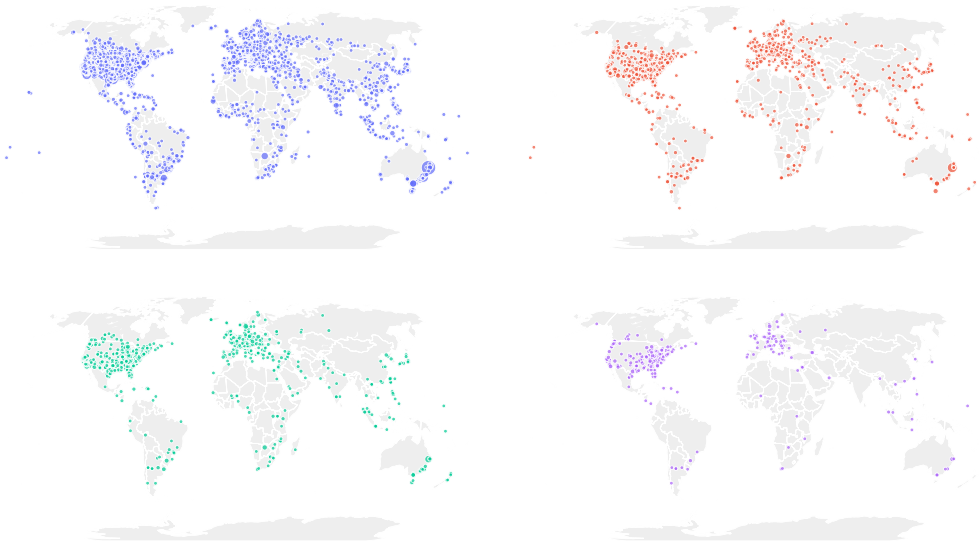

Fig. 4. Geographic distribution of road users represented in the crash corpus. The four panels show the most prevalent road user categories: top left, car; top right, truck; bottom left, van / pickup truck; and bottom right, motorcycle. Across all 34,295 accepted crash segments, cars occur in 32,213 segments (93.93%), trucks in 6,596 (19.23%), vans / pickup trucks in 3,369 (9.82%), and motorcycles in 1,186 (3.46%).

segments recorded in the persistent state. The published latitude and longitude values correspond to canonical locality reference coordinates rather than the exact position of the recorded event.

The walking and crash corpora therefore exercise different parts of the same underlying infrastructure. The walking corpus consists primarily of long form observations of pedestrian movement across everyday environments, whereas the crash corpus contains large numbers of shorter safety critical events extracted from heterogeneous source videos. Despite these differences, both corpora retain structured semantic annotations, explicit geographic evidence and provenance, a versioned processing state, and reproducible dataset outputs.

4 Discussion

This study demonstrates that public video platforms can support mobility datasets at a scale and contextual diversity that would be difficult to reproduce through purpose built recording alone, provided that collection is treated as an infrastructure problem rather than as a sequence of downloads. The two resulting corpora illustrate complementary uses of this infrastructure. The walking corpus captures long form observations of everyday pedestrian movement across streets, markets, parks, waterfronts, transport hubs, and other public environments, while the crash corpus captures shorter and comparatively rare safety critical events. Together, they show that the same underlying approach can support substantially different mobility research questions while preserving a common structure for discovery,

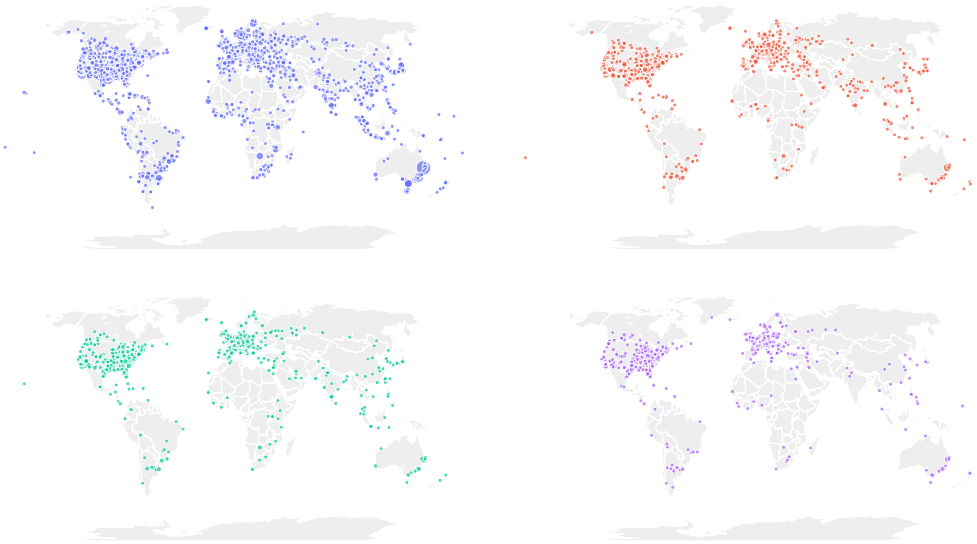

Fig. 5. Geographic distribution of crash segments by MMUCC manner of collision. The four displayed panels show the most frequent collision manners in the synthetic completion used for this draft figure: top left, front to rear or rear to front collision, 10,793 segments (61.59%); top right, angle collision, 2,559 (14.60%); bottom left, same direction sideswipe, 1,462 (8.34%); and bottom right, front to front collision, 995 (5.68%).

screening, temporal segmentation, semantic annotation, geographic grounding, provenance, and reproducible dataset releases.

Public video should therefore be viewed as complementary to, rather than a replacement for, purpose built mobility datasets such as KITTI, nuScenes, Waymo, or dedicated pedestrian and crash benchmarks. Controlled datasets provide consistent sensing configurations, calibration, and carefully designed acquisition procedures, whereas public video broadens the observable range of places, recording devices, viewpoints, behaviours, and event types. This complements geographically diverse continuous driving resources such as CROWD [2]. When observations of everyday driving, pedestrian activity, infrastructure, and safety critical events are considered together, they can provide increasingly rich representations of real world mobility environments and form a useful data foundation for future world creation and simulation. This is particularly relevant to human centred mobility research, where questions concerning pedestrian experience, interaction with infrastructure, road user encounters, and unusual traffic events often depend on naturalistic context rather than on a tightly controlled sensor configuration. At the same time, public video reflects who records, what they choose to upload, and what the platform makes discoverable, and the resulting corpora should therefore not be interpreted as representative samples of global mobility behaviour.

The differences between the walking and crash corpora illustrate how the characteristics of the source material influence what can be extracted reliably. The walking corpus retained 201.29 hours of footage and provided strong geographic coverage, with 90.23% of accepted segments resolving to a locality across 28 countries or territories. Its long

form videos also supported descriptions of walking environment, time of day, and geographic evidence obtained from video metadata, uploader supplied timestamps, and text visible within the footage. The crash corpus retained 151.35 hours of traffic safety footage and contained a much larger number of individual segments, but geographic resolution was possible for 51.28% of accepted segments. Crash compilations are frequently assembled from recordings originating from different places, and the individual clips may be detached from the descriptive context of the original recording. At the same time, they provide access to visually observable road user involvement, collisions, and near collisions that would be difficult to collect prospectively at a comparable scale. These differences support the use of corpus specific annotations while retaining a shared underlying data infrastructure.

A recurring design choice throughout both pipelines was to preserve uncertainty rather than force every segment into a complete record. This is particularly important for geographic information. A candidate place can remain recorded even when the available evidence is insufficient to publish it as a resolved locality. In some videos, the latitude and longitude values are directly visible as overlays or subtitles, including the coordinates automatically added by some dashcams. Similar examples were encountered in the continuous driving videos collected for CROWD [2]. Such coordinates provide valuable direct geographic evidence, but their availability and precision cannot be assumed consistently across heterogeneous public videos. We therefore use them to identify and canonicalise the containing locality rather than presenting them as exact event positions. Roads, stations, businesses, counties, and other points of interest are likewise not silently substituted for cities or towns. The same principle applies to semantic annotation. For example, the crash corpus records visually supported event characteristics but does not infer fault, impairment, injury severity, or other high-contribution attributes from indirect visual cues. This reduces the number of complete labels, but makes their meaning and provenance clearer. For researchers using the resulting datasets, preserving the distinction between candidate evidence, resolved output, and unknown values also makes it possible to construct stricter or more permissive subsets without discarding the original machine generated evidence.

The results also illustrate the practical value of treating the dataset as an evolving process rather than as a static collection. Public platform sources change over time: new videos appear, older videos disappear, metadata may change, and annotation procedures can improve. By recording how each video has already been processed, the framework can update only the parts affected by a methodological change. For example, an improved geographic procedure can be applied again without repeating video discovery or temporal segmentation, while a revised crash taxonomy can be added only to segments that still require the new annotations. Previously completed work that remains valid can, therefore, be reused. This separation between an evolving collection and fixed dataset releases is important for reproducibility: researchers can cite the exact snapshot used in a study while the underlying corpus continues to grow and improve.

Although recent general purpose AI models can perform many of the tasks required here, using one model for the entire pipeline is neither necessary nor always desirable. Different stages operate on different kinds of information and have different computational requirements. Metadata screening is primarily a text based task and can be performed efficiently without analysing the full video, whereas deciding whether a segment contains genuine walking, a crash or near collision, or explicit geographic evidence requires direct inspection of visual and temporal information. Other operations, such as detecting candidate editing boundaries, downloading videos, maintaining processing history, and resolving place names, are better handled by conventional software or dedicated services rather than by a generative model. We therefore use different components according to the information required at each stage: Qwen performs the initial metadata screening, while Cosmos is used when decisions depend on evidence contained within the video. This modular design also makes the process easier to inspect and update because one component can be replaced or improved without requiring the complete dataset to be reconstructed. More generally, increasingly capable models do

not remove the need to decide which parts of dataset construction should be delegated to AI, what evidence a model should be allowed to use, and where simpler and more reproducible computational procedures are preferable.

The use of automated models also has implications for how the resulting annotations should be interpreted. Structured outputs, validation rules, confidence requirements, and repeated attempts can reduce obvious failures, but they do not turn model generated annotations into manually verified ground truth. Models may still miss relevant footage, misread text, or assign an incorrect contextual category, and these errors can propagate if the annotations are subsequently used to train or evaluate other systems. Manual verification therefore remains appropriate when a study depends on precise event categories, fine geographic distinctions, rare subclasses, or other labels for which annotation errors could materially affect the conclusions. Public availability also does not remove privacy, copyright, or ethical considerations, particularly when visual observations are combined with geographic information.

A natural next step is to move from video level or segment level geographic grounding towards finer spatial annotation of the visual stream. Where sufficient evidence is available, individual frames could be associated with geographic coordinates rather than assigning only a canonical locality to an entire segment. Combining frame level geographic information with the existing semantic annotations would make it possible to construct spatial heatmaps of pedestrian activity, traffic behaviour, road user interactions, and safety critical events. Such representations could support analyses of how mobility behaviour changes within and across cities rather than only how often particular behaviours appear in a collection. In combination with larger resources such as CROWD and other sources of road structure, infrastructure, and dynamic road user observations, this could also contribute towards geographically grounded representations for world creation, simulation, and world model research.

Overall, the contribution of this work is not a particular search query, annotation model, or individual corpus, but a reusable approach for transforming heterogeneous public video into structured, traceable, and continuously maintainable mobility data. The walking and crash use cases demonstrate that the same infrastructure can support both long form observations of everyday mobility and large collections of rare safety critical events while retaining explicit evidence and uncertainty throughout the process. Public video cannot provide the sensing consistency or sampling control of purpose built datasets, but it can substantially broaden the geographic, behavioural, and contextual range available to mobility researchers. Treating these sources as maintained research infrastructure rather than disposable downloads makes that breadth usable while preserving the transparency and reproducibility required for scientific analysis.

5 Limitations and future studies

Public video platforms do not provide a representative sample of mobility. This study focuses on YouTube because it is currently the most practical large scale source for this type of public video, but similar biases are likely to affect other platforms. Coverage depends on platform access, uploader behaviour, language, tourism, camera ownership, internet connectivity, search ranking, and the chosen discovery queries. Wealthier countries are also likely to contribute more footage. Corpus frequencies should therefore not be interpreted as population prevalence, exposure, crash risk, or the true frequency of a behaviour or environment.

The semantic annotations are automatically generated and may contain errors. Models can miss relevant footage, retain unsuitable footage, misread visible text, or assign an incorrect category. The labels should therefore not be treated as manually verified ground truth, and studies requiring precise event categories, environmental attributes, or rare subclasses should validate the relevant subset independently.

Geographic information is intentionally conservative. Many videos contain no explicit locality evidence, and the published coordinates represent canonical locality reference points rather than exact camera or event positions. This limits fine scale spatial analysis but avoids implying unsupported precision. Future work could investigate frame level geocoding where reliable geographic evidence is available and use these locations to create spatial heatmaps of pedestrian activity, traffic behaviour, and safety critical events.

Public videos may also be removed, made private, or become region restricted after a dataset release. Source identifiers and derived annotations preserve provenance but cannot guarantee continued access to the original media. Ethical, privacy, and copyright considerations also remain important, particularly because mobility videos may contain identifiable people, vehicles, homes, workplaces, and other sensitive contextual information.

The framework also depends on changing platforms, models, geocoding services, and software. Fixed dataset snapshots and recorded processing versions improve reproducibility, but cannot completely remove these dependencies. The approach itself is not limited to YouTube, and future studies could extend it to other public video platforms to broaden geographic and behavioural coverage and reduce dependence on a single source.

6 Supplementary material

In line with current open science practices and recommendations for transparency in automotive user research [10], the authors openly provide these research artefacts to support reproducibility, collaboration and further advancements in the field. The generated videos, their description, and analysis code are available at https://doi.org/10.4121/3ee08a97-f065-4811-a6c8-aeb4191ae102. The maintained version of the code for car crash is available at https://github.com/Shaadalam9/car-crash-yt and for walking is at https://github.com/Shaadalam9/walking-yt.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675* (2016). doi:10.48550/arXiv.1609.08675

[2] Md Shadab Alam, Olena Bazilinska, and Pavlo Bazilinskyy. 2026. *A global dataset of continuous urban dashcam driving.* Technical Report. arXiv:2604.01044 [cs.CV] https://arxiv.org/abs/2604.01044

[3] Md Shadab Alam and Pavlo Bazilinskyy. 2026. Nineteen Years of ASMR on YouTube: A Multilingual, Theme-Level Analysis of 72,070 Videos. (2026).

[4] Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-Based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In *Proceedings of the 28th ACM International Conference on Multimedia.* Association for Computing Machinery, 2682–2690. doi:10.1145/3394171.3413827

[5] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. doi:10.1162/tacl_a_00041

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 11621–11631. doi:10.1109/CVPR42600.2020.01164

[7] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23).* Curran Associates Inc., Red Hook, NY, USA, Article 379, 12 pages. doi:10.48550/arXiv.2309.16020

[8] Claire Chambers, Gaiqing Kong, Kunlin Wei, and Konrad Kording. 2019. Pose Estimates from Online Videos Show That Side-by-Side Walkers Synchronize Movement under Naturalistic Conditions. *PLOS ONE* 14, 6 (2019), e0217861. doi:10.1371/journal.pone.0217861

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3213–3223. doi:10.1109/CVPR.2016.350

[10] Patrick Ebel, Pavlo Bazilinskyy, Mark Colley, Courtney Michael Goodridge, Philipp Hock, Christian P. Janssen, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, and Philipp Wintersberger. 2024. Changing Lanes Toward Open Science: Openness and Transparency in Automotive User Research. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Stanford, CA, USA) *(AutomotiveUI '24).* Association for Computing Machinery, New York, NY, USA, 94–105. doi:10.1145/3640792.3675730

[11] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. 2019. DADA-2000: Can Driving Accident Be Predicted by Driver Attention? Analyzed by a Benchmark. In *2019 IEEE Intelligent Transportation Systems Conference.* IEEE, 4303–4309. doi:10.1109/ITSC.2019.8917218

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daum'e III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361. doi:10.1109/CVPR.2012.6248074

[14] Maximilian C. Hartmann and Ross S. Purves. 2023. Seeing through a New Lens: Exploring the Potential of City Walking Tour Videos for Urban Analytics. *International Journal of Digital Earth* 16, 1 (2023), 2555–2573. doi:10.1080/17538947.2023.2230182

[15] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4110–4124. doi:10.18653/v1/2021.naacl-main.324

[16] Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. 2023. CORE: A Global Aggregation Service for Open Access Papers. *Scientific Data* 10 (2023), 366. doi:10.1038/s41597-023-02208-w

[17] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A Continuously Growing Dataset of Sentential Paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1224–1234. doi:10.18653/v1/D17-1126

[18] Ana-Maria Marcu, Long Chen, Jan H"unermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. 2024. LingoQA: Visual Question Answering for Autonomous Driving. In *Computer Vision – ECCV 2024*. Springer, 252–269. doi:10.1007/978-3-031-72980-5_15

[19] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640. doi:10.1109/ICCV.2019.00272

[20] Daniel Moura, Shizhan Zhu, and Orly Zvitia. 2025. Nexar Dashcam Collision Prediction Dataset and Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2608–2616. doi:10.48550/arXiv.2503.03848

[21] National Highway Traffic Safety Administration. 2025. *MMUCC Guideline: Model Minimum Uniform Crash Criteria, Sixth Edition*. Technical Report DOT HS 813 525. National Highway Traffic Safety Administration, U.S. Department of Transportation. Revised edition; MMUCC Sixth Edition updated in 2024.

[22] Sara Nielsen, Mikael B. Skov, Karl Damkjær Hansen, and Aleksandra Kaszowska. 2023. Using User-Generated YouTube Videos to Understand Unguided Interactions with Robots in Public Places. *ACM Transactions on Human-Robot Interaction* 12, 1, Article 5 (2023). doi:10.1145/3550280

[23] Nominatim Development Team. 2026. Nominatim Manual. OpenStreetMap geocoding and reverse geocoding documentation.

[24] NVIDIA. 2026. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800* (2026). arXiv:2606.02800 https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf

[25] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 1776–1826. doi:10.1145/3531146.3533231

[26] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL]

[27] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6262–6271. doi:10.1109/ICCV.2019.00636

[28] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. DriveLM: Driving with Graph Visual Question Answering. In *Computer Vision – ECCV 2024*. Springer, 256–274. doi:10.1007/978-3-031-72943-0_15

[29] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2446–2454. doi:10.48550/arXiv.1912.04838

[30] World Health Organization. 2023. *Global Status Report on Road Safety 2023*. World Health Organization, Geneva, Switzerland. https://www.who.int/publications/i/item/9789240086517

[31] Zimin Xia and Alexandre Alahi. 2025. FG^2: Fine-Grained Cross-View Localization by Fine-Grained Feature Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6362–6372. doi:10.48550/arXiv.2503.18725

[32] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2636–2645. doi:10.48550/arXiv.1805.04687

[33] Xinyan Yu, Marius Hoggenmüller, Tram Thi Minh Tran, Yiyuan Wang, and Martin Tomitsch. 2024. Understanding the Interaction between Delivery Robots and Other Road and Sidewalk Users: A Study of User-Generated Online Videos. *ACM Transactions on Human-Robot Interaction* 13, 4, Article 59 (2024), 32 pages. doi:10.1145/3677615