ICS 43.020
CCS R85

团 体 标 准

T/SHV2X 11—2025

汽车驾驶自动化系统通用语料库 第4部分：语料数据清洗

General corpus for automotive driving automation system—
Part 4: Data cleaning

2025－07－25 发布　　2025－07－25 实施

上海市车联网协会　　发 布

目　　次

前　言

《汽车驾驶自动化系统通用语料库》系列标准拟分为六个部分：

——第1部分：总体要求；

——第2部分：术语和定义；

——第3部分：语料数据采集；

——第4部分：语料数据清洗；

——第5部分：语料数据标注；

——第6部分：语料数据测试。

本文件为第4部分。

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市车联网协会提出并归口。

本文件起草单位：上海库帕思科技有限公司、智能汽车创新发展平台（上海）有限公司、智己汽车科技有限公司、地平线征程（上海）科技有限公司、上海临港绝影智能科技有限公司、上海机动车检测认证技术研究中心有限公司、上海金桥智能网联汽车发展有限公司、北京一辅智行科技有限公司、上海优咔网络科技有限公司、亿咖通（上海）技术有限公司、滴水智行科技有限公司、上海航盛实业有限公司、上海阶跃星辰智能科技有限公司、东华大学、上海交通大学、上研智联智能出行科技（上海）有限公司、北京赛目科技股份有限公司、苏州柏川数据科技有限公司、华为技术有限公司。

本文件主要起草人：山栋明、黄海清、施佳樑、郭辉、赵九花、贺锦鹏、蒋达夫、周剑鸣、黄剑其、邵亚萌、徐春雷、解瀚光、朱雷、谭龙欢、于峰、滕添益、张帆、林瑜、周轶、袁月明、李学根、张裕珍、曹宇、邓思文、贺仁驹、李晨歌、马昊、陈紫娟、杨闻博、丁楚晨、李想、蔡雨辰、黄鹏飞、刘建业、李勋宏、马骏、李轶刚、刘壹青、田浩、郭晓宾、董连飞、范昌琪、李璟、孙雯、陈巧慧、王娜、沈滨、孔令和、鲁江东、汪大明、徐鹏、何丰、谭哲、薛晓卿、刘兴、马东升、刘鹏宇、邓子涵。

本文件首批承诺执行单位：上海库帕思科技有限公司、智能汽车创新发展平台（上海）有限公司、智己汽车科技有限公司、地平线征程（上海）科技有限公司、上海临港绝影智能科技有限公司、上海机动车检测认证技术研究中心有限公司、上海金桥智能网联汽车发展有限公司、北京一辅智行科技有限公司、上海优咔网络科技有限公司、亿咖通（上海）技术有限公司、滴水智行科技有限公司、上海航盛实业有限公司、上海阶跃星辰智能科技有限公司、东华大学、上海交通大学、上研智联智能出行科技（上海）有限公司、北京赛目科技股份有限公司、苏州柏川数据科技有限公司、华为技术有限公司。

引　言

为了实现可靠的汽车驾驶自动化，大量准确的语料数据至关重要。语料数据是自动驾驶系统的基础，它能够帮助系统理解复杂的交通环境、做出明智的决策，并不断优化性能。随着端到端模型整合度越来越高，对训练语料的规模和质量要求也成倍增加。大规模、高质量的数据标注（特别是端到端感知数据和推理数据）是构建安全、可靠、高性能自动驾驶系统的基石。围绕“采、洗、标、测”一体化流程，制定《汽车驾驶自动化系统通用语料库》系列标准，包括：

——第1部分：总体要求，明确系列标准建设的总体要求，为后续各部分标准的制定提供指导和基础；

——第2部分：术语和定义，统一系列标准建设过程中涉及的关键术语和定义；

——第3部分：语料数据采集，对语料数据资源格式提出规范要求；

——第4部分：语料数据清洗，针对采集好的数据，对语料数据清洗的流程与质量要求提出规范要求；

——第5部分：语料数据标注，针对清洗好的数据，对语料数据标注的适用场景、标注内容、标注方式、数据存储提出规范要求；

——第6部分：语料数据测试，针对标注好的数据，对语料数据测试的流程和质量要求提出规范要求。

本文件为第4部分语料数据清洗。通过本文件的制定，对自动驾驶数据清洗流程与质量进行统一规范，保障感知系统训练过程中的可靠性，最终可生成高价值的语料数据集用于模型训练与决策，为自动驾驶技术的发展提供有力支持。通过构建可复用的标准化自动驾驶训练数据集，促进汽车驾驶自动化系统语料资源高效流通利用，降低业内企业在重复数据采集、清洗、标注等方面的巨额成本，提升汽车企业竞争力且推进汽车产业健康发展。

汽车驾驶自动化系统通用语料库 第4部分：语料数据清洗

1 范围

本文件规定了汽车驾驶自动化系统通用语料库语料数据清洗的流程，包括数据预处理、数据清洗和质量等方面的要求。

本文件适用于企业、研究机构对汽车驾驶自动化系统通用语料库的研究、开发、维护、应用、评估等工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41871—2022 信息安全技术 汽车数据处理安全要求

CH/T 8023—2011 机械激光雷达数据处理技术规范

3 术语和定义

T/SHV2X 10—2025文件界定的术语和定义适用于本文件。

4 符号和缩略语

下列缩略语适用于本文件。

IMU 惯性测量单元（Inertial Measurement Unit）

5 数据预处理

5.1 数据格式统一

数据格式统一，指在自动驾驶语料数据处理过程中，待清洗的所有原始采集数据须转换为统一、可解析、可追溯的结构化格式数据集，相关数据资源的要求应满足T/SHV2X 1—2025的规定。

5.2 完整性校验

在完成数据格式统一后，数据资源应按表1中的要求和规则进行数据包完整性、元数据完整性及数据融合和对齐等类型的校验和问题处理。

表1 完整性校验及问题处理规则

类型	要求	问题处理规则
数据包完整性	采用哈希校验值与数据采集记录进行比对	如发现缺失情况，应触发数据重传/作废流程
元数据完整性	对传感器个数、传感器内外参数文件、场景标签文件等进行检查	如发现缺失情况，应触发数据重传/作废流程
数据融合和对齐	对多传感器数据进行时间戳同步检查、坐标系融合对齐检查	如发现缺失和误差较大情况，应触发数据作废/修复流程

5.3 数据抽帧

在完成完整性校验后，数据资源应按表2中的要求和规则进行数据抽帧处理。

表 2 数据抽帧要求及问题处理规则

类型	要求	问题处理规则
低速/静态场景	对车辆拥堵、红灯时间长等高冗余场景选择性丢弃冗余数据帧	按照固定间隔抽帧
结构化道路巡航场景	对高速公路直线行驶等低交互复杂度场景选择性丢弃冗余数据帧	按照时间均匀采样抽帧
动态交互复杂场景	对高交互复杂场景进行数据保留	不建议抽帧，保留完整连续帧序列
行为预测和轨迹建模场景	对车辆切入变道、行人横穿马路等关键事件场景进行数据保留	不建议抽帧，保留完整连续帧序列

6 数据清洗

在完成数据抽帧后，数据资源应按表3中的要求和规则进行噪声数据、重复数据、静止数据、无效/低价值数据等类型的清洗和问题处理，以及高价值数据的筛选。

表 3 数据清洗要求及问题处理规则

类型	要求	问题处理规则
噪声数据	对图像/视频数据，检查画面亮度失衡、模糊/拖影、遮挡/花屏、色偏等内容	如发现缺失情况，应触发数据删除流程
	对雷达数据，检查离群点、点云缺失、点云密度异常等内容	
	对IMU数据，检查信号丢失，速度或加速度异常造成的空值数据等内容	
重复数据	检查重复采集的无效帧或高度相似数据	如发现重复情况，应触发数据删除/保留首帧流程
静止数据	检查在车辆通过的时间尺度内，长时间（秒至分钟级）保持几何形态与位置不变产生的数据	如发现静止情况，应触发数据删除/保留首帧流程
其他无效/低价值数据	检查数据中的无效信息	如发现低价值情况，应触发数据删除流程
高价值数据筛选	检查极端情况数据	如选取稀有场景数据、高复杂度场景数据

7 质量要求

在自动驾驶语料数据处理过程中，需通过技术手段识别、修正或移除数据集中存在的错误、噪声、不一致、冗余或无效内容，提升数据质量。

汽车驾驶自动化系统通用语料库的语料数据质量检验应包含：

a） 完整性指标，如有效数据完整率、有效时间覆盖率等，其中完整率宜≥95%，覆盖率宜≥90%；

b） 准确性指标，如传感器数据准确率，异常值宜≤5%；

c） 一致性指标，如多传感器时间同步精度宜≤10 ms，同类传感器测量结果一致性偏差宜≤5%；

d） 符合GB/T 41871—2022中数据处理安全要求。

8 数据存储

数据存储应按照以数据批次为单位进行管理，单位批次内应包含完整传感器配置参数、标定文件及原始数据。原始数据文件应逐帧以系统时间戳（精确至毫秒）命名。结构化数据包(标准目录树+索引文件) 样例文件结构参见附录A。

附 录 A
（资料性）
文件结构样例

A.1 结构化数据包样例文件结构

自动驾驶数据清洗中，常见数据包文件结构样例如下：

```
|raw_data_+时间(如：2025_03_01_13_00，即2025年03月01日13点00分)
    |—info
        |—basic_info.json
        |—environment_info.json
        |—vehicle_info.json
        |—task_info.json
        |—calibration_info.json
    |—trajectory
        |—trajectory.json
    |—camera_x(x为相机编号)
        |—pics
        |—timestamp00.jpg
        |—timestamp01.jpg
    |—lidar_y(y为激光编号)
        |—pcd
        |—timestamp00.pcd
        |—timestamp01.pcd
```

参 考 文 献

[1] ISO8601:2019 数据存储和交换形式·信息交换·日期和时间的表示方法
[2] T/SAIAS 015—2024 语料库建设导则
[3] T/SHV2X 9—2025 汽车驾驶自动化系统通用语料库 第1部分：总体要求
[4] T/SHV2X 10—2025 汽车驾驶自动化系统通用语料库 第2部分：术语和定义
[5] T/SHV2X 1—2025 汽车驾驶自动化系统通用语料库 第3部分：语料数据采集
[6] 智能网联汽车时空数据安全处理基本要求
[7] 自然资发〔2024〕139号 自然资源部关于加强智能网联汽车有关测绘地理信息安全管理的通知