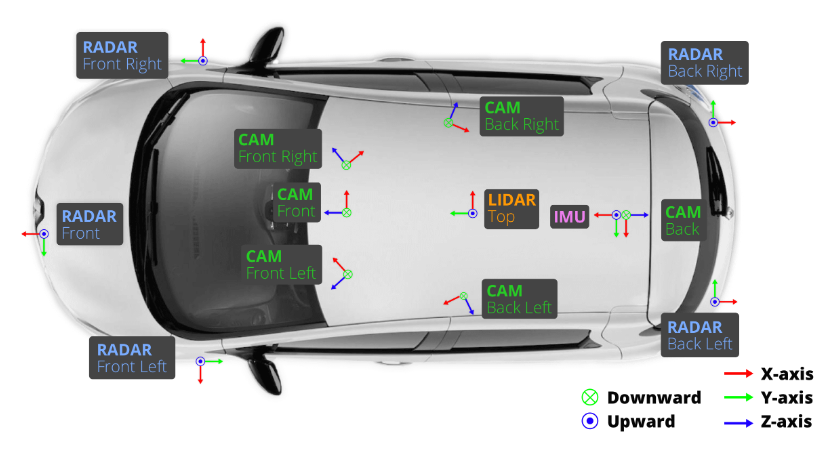
NuScenes –

* 1000 driving scenes in Boston and Singapore of 20s length.
* It has 23 object classes with 3D bounding boxes at 2Hz over the entire dataset.
* Semantic maps, baseline routes, ego\_pose derived from lidar data.
* It provides data from the entire sensor suite of an autonomous vehicle (6 cameras, 1 LIDAR, 5 RADAR, GPS, IMU, CAN data).
* Driving routes have a diverse set of locations, times and weather conditions.
* To balance the class frequency distribution, more scenes with rare classes are included (such as bicycles).
* 
* Inputs:
* 1 spinning LIDAR (Velodyne HDL32E):   
  20Hz, 32 beams spinning lidar, 360 horizontal and +10 to -30 vertical FOV, 80-100m range.
* 5 long RADAR sensor (Continental ARS 408-21):  
  <=250m range, 13Hz, measures dis and velocity in one cycle using FMCW
* 6x camera (Basler acA1600-60gc):  
  12Hz, 1/1.8'' CMOS sensor of 1600x1200 resolution – cropped to 1600x900 roi to reduce processing, images unpacked to rgb format and compressed to jpeg, auto exposure.
* 1 IMU and GPS:  
  position accuracy of 20mm

nuScenes is token/JSON-based. Below are the JSON tables / files that form a nuScenes database:

1. sensor - list of sensor definitions (camera / lidar / radar) and sensor channels (name, modality, etc.).

2. calibrated\_sensor - per-sensor intrinsics & extrinsic

3. ego\_pose - vehicle pose in global coordinates at timestamps (translation, rotation, timestamp).

4. log - data about each data log / drive (date, location, vehicle, raw log file).

5. scene - A scene is a 20s long sequence of consecutive frames.

6. sample - nuScenes annotation keyframes are at 2 Hz.

7. sample\_data.json - entries for every sensor frame. Allows dense-sweep frames plus annotated keyframes.

8. sample\_annotation - the 3D bounding boxes for each sample.

9. instance - instance tracking data enabling tracking across frames within a scene.

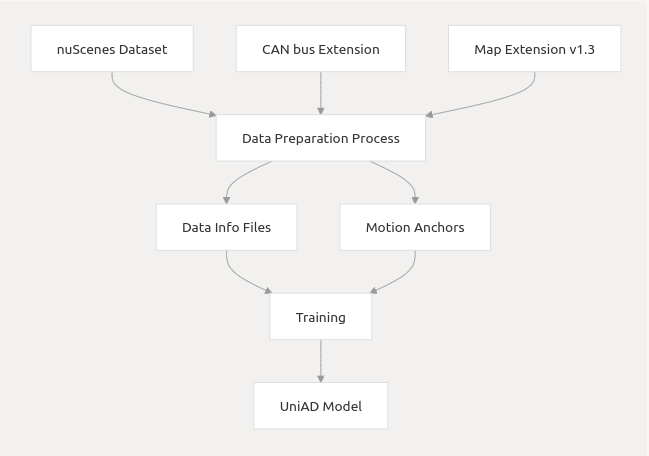
10. category & attribute - category taxonomy (e.g., vehicle.car, human.pedestrian) and attributes (moving/stationary, etc.).

11. map - nuScenes stores semantic map masks (top-down drivable area)

12. visibility - optional object visibility labels (occlusion).

NuScenes used in Uniad, GenAD, VAD/VADv2, etc

NuScenes in UniAD:



UniAD is a planning oriented multitask BEV model. It is trained on nuScenes benchmark. NuScenes is used for lot of tasks like Detection, Tracking, Mapping, Motion Prediction, Planning. It takes:  
nuScenes V1.0 full dataset: Contains camera images, LiDAR point clouds, and annotations  
CAN bus extension: Provides vehicle state information including pose, velocity, and acceleration  
Map extension (v1.3): Contains HD map information for the scenes  
  
JSON files provided.



Inputs:  
Multi view camera images, lidar, calibration + ego pose, 3D annotations(with IDs), Map layers, can  
  
key files: scene(sequence of samples), sample(snapshot), sample data(sensor files), ego pose(car’s pose at that timestamp), calibrated sensor(intrinsics/extrinsics of each sensor relative to ego vehicle), annotation, map(HD map layers).

Background - Unified autonomous driving – Unified full stack tasks -detection, tracking, segmentation, planning – better info sharing, coordination and safety – used nuScenes dataset   
Process – Planning oriented ad. A query based multitask design (Transformer) - Instead of separate modules, tasks communicate via a unified query interface, and all tasks attend to a common global representation. All perception and prediction modules are designed in a transformer decoder structure, with task queries as interfaces connecting each node.  
It consists of 6 specialized modules – i) Backbone -processes images from multiple cameras and transforms them into a unified BEV format  
  
ii) TrackFormer -based on multi object tracking research (MOTR), allows for continuous object tracking with consistent IDs for each object across frames. It consists of two queries – detection query (to detect new objects that appear) and track queries (to follow already detected objects). And queries share information across time using self-attention module. It also has an ego vehicle query to represent the self-driving vehicle.  
  
iii) MapFormer -real time mapping by identifying road features like lanes, dividers, etc. by segmenting env into diff things enabling UniAD to adapt to road conditions without high-definition maps. Its design is based on Panoptic SegFormer. Outputs map queries.  
  
iv) Motion Former –takes queries form TrackForner and MapFormer to predicts future trajectories (top k trajectories) of agents enabling UniAD to account for uncertainties. It works in one pass which makes it fast compared to methods that predict separately for each agent. It captures 3 types of interactions – agent- agent, agent-map, agent-goal point.  
  
v) OccFormer -creates a forecasted occupancy map in grid map format predicting which areas will be occupied by other vehicles/pedestrians, essential for proactive collision avoidance. Earlier RNN methods were used, but Occformer instead directly links agents + scene-level features to predict occupancy.  
  
vi) Planner - creates a safe and efficient trajectory by integrating from previous modules.  
Training pipeline – stage 1(perception training) and stage 2 (end to end training)

GenAD inputs –   
Multi view camera images, ego pose, map, calibration  
VAD Inputs-  
Camera images, vectorized scene representation, ego pose, calibration

All related Indian datasets Comparison:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | nuScenes | IDD-3D | DriveIndia | TIAND(TiHAN) | IDD | IDD Multimodal |
| Multi-camera | 6 cameras | 6 cameras | Front camera only | 4 cameras | Single camera | Stereo(front) |
| LiDAR | Yes | Yes(12k lidar frames) | No | Yes | No | Y |
| Radar (not used) | Yes | No | No | Yes (6 radars) | No | N |
| Ego pose (GPS/IMU) | Yes | Na | No | Yes | No | GPS at 15hz |
| 3D boxes | Yes | Yes | No | NA (maybe 2D boxes) | No | Not sure |
| Tracking IDs | Yes | Yes | N | NA | N | Not sure |
| Intrinsics/ extrinsics | Y | Y | N | Y | N | Intrinsics only |
| Map | Yes | No | No | No | No | No |
| CAN bus data | Yes (20hz) | N | N | N | N | Yes(OBD data) |
| Comments |  | Strong for perception+tracking; still needs map/can |  | Map and can missing |  | Labels n map missing; great ego state and depth |

Indian Datasets:

More traffic density, heterogeneity, road infrastructure, weather and lighting

1. IDD -It consists of 10,000 images, finely annotated with 34 classes collected from 182 drive sequences on Indian roads.  
   Hyderabad, Bangalore cities

No lidar, no tracking ids, single frame

1. IDD-3D  
   6 cameras (10-15Hz) rgb images and pcd from 1 lidar(10Hz)  
   contains 15.5k synchronized frames of which 12k lidar frames are annotated  
   Hyderabad  
   lidar frames annotated with 3D bounding box around objs labelled with 17 classes.  
   Sequence of 100 frames at 10fps making 150 sequences, each of 10s.  
   3D bounding box annotations – 17 categories  
   Saved in ros-bag format
2. DriveIndia   
   only images, 2D object detection dataset - 66986 high res images, 24 India-specific classes)  
   maybe good for fine tuning cameras?
3. IDD multi-modal   
   stereo images (front camera pair) at 15fps  
   GPS points  
   16 channel Lidar  
   OBD data

Stereo images, LIDAR data, GPS & CAN bus info  
Supports sensor fusion, localization, and 3D scene reconstruction.

1. TIAND (TiHAN-IIT hyderabad)   
   4 camera feeds (1920x1200 @42 fps, 42Hz)  
   6 RADARs (1long range and 5 short range units, ~20Hz)  
   1 LIDAR (360degree 128 channel, 10-20 Hz)  
   GNSS and data for sensors (ego pose, etc)  
   150 scenes of 2-4 minute scenes  
   Hyderabad, India  
   3D OD for lidar and 2D OD for camera  
   Exact frame count not stated   
   Can access by requesting through gform.  
     
   -Can error
2. DIRS24.v1 and DIRS21– iitr   
   12- megapixel camera with an f/1.8 aperture with 1920 X 1080 pixels resolution.   
   5093 images were collected under various scenes, weather, season, and illumination conditions.   
   seven classes of objects: Pedestrian, Rider, Car, Bus, Truck, NMV (Non-Motorized Vehicle), and EA (Electric Auto).
3. Indian Traffic Videos (MP4)  
   500 hours of video only  
     
   not much useful as only images not video or multimodal sensors.
4. IDD-X  
   videos with text annotations  
   dual-view dataset for analyzing driving behavior in complex traffic, with rich object annotations and explainability labels.  
   3634 annotated driving scenarios in 1140 untrimmed videos. The annotations include 697K important object bounding boxes (9K object tracks), 1-12 objects per driving scenario covering 10 categories, and 19 explanation label categories.

To make IDD- Multimodal dataset similar to nuScenes:  
  
i) Has diff sensor config ?– stereo front facing camera (nuscenes-6x cameras), 16 channel lidar (nuscenes-32channel lidar), GPS

ii) Calibration adjustment – No data on calibration info.   
Can try to extract intrinsic and extrinsic info from IDDM   
If not provided, can align them with respect to camera and lidar. Intrinsics wrt camera specs and extrinsic for camera-to-vehicle and lidar-to-vehicle

iii) Timestamp synchronization?- will have to align lidar and camera input at the same time. Camera and GPS both at 15hz, LiDAR freq not sure (nuscenes – camera 12hz, lidar 20hz)

iv) Map – Can reconstruct local 3D maps from camera footage - (DashIndo3D)  
 or can try skipping.

v) Conversion of IDDM logs into JSON format containing sample, sensor calibration, ego\_pose, annotations.

vi) Annotations - Will need to generate 3D labels

vii)Ego pose – only gps (latitude/longitude) given at 15hz. Orientation or full pose is not given.