KPIT

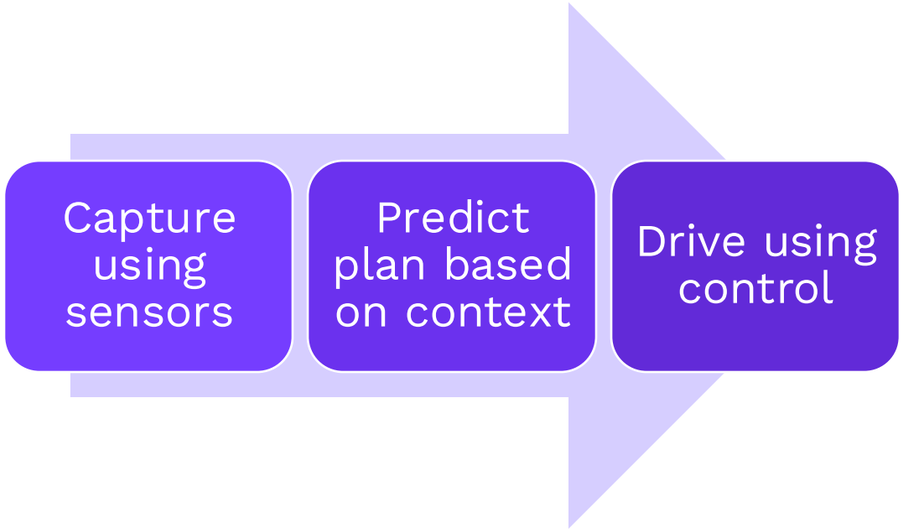
25/7/2025

E2E Stack Design Flow – CXO Level Overview

RaghuRam Theerthala

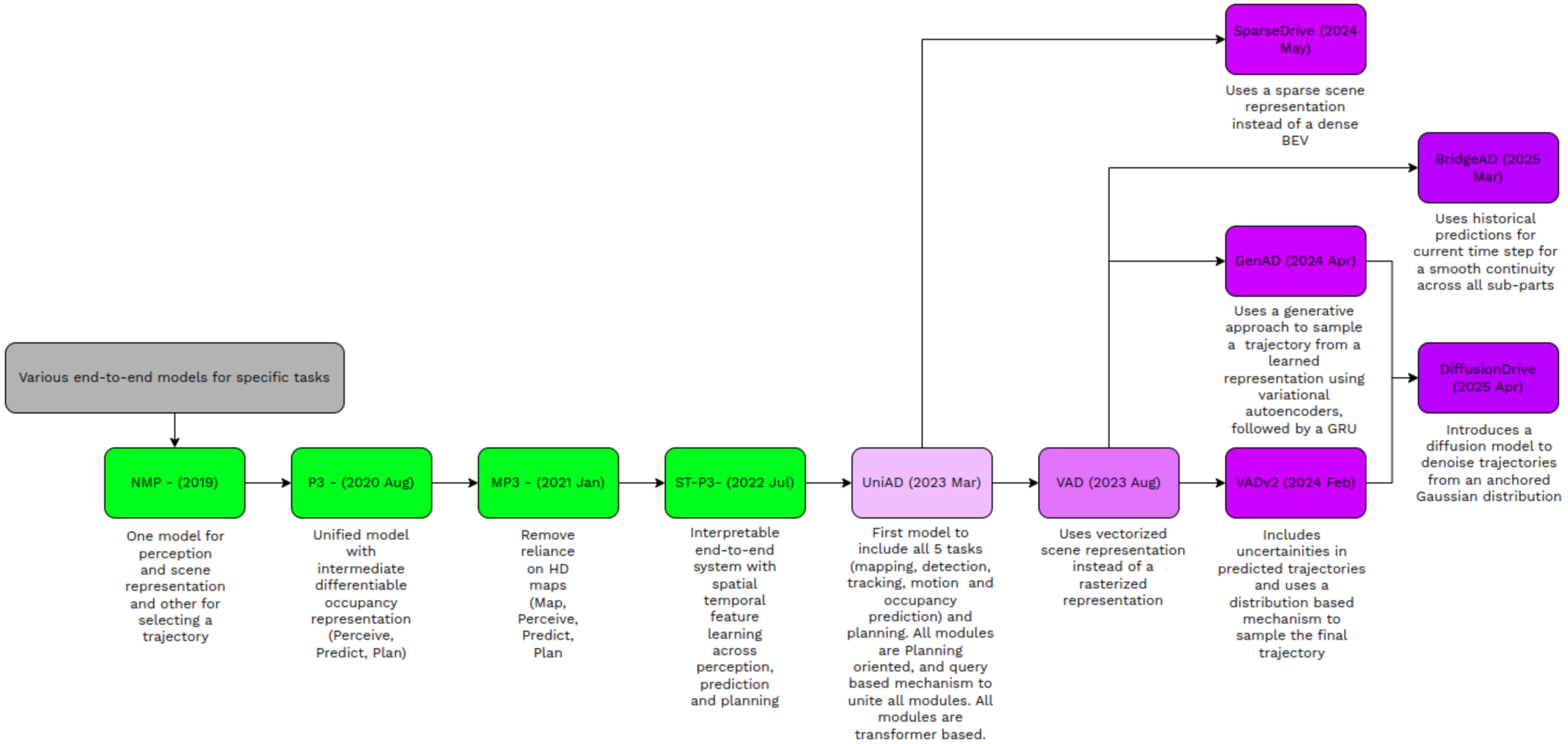
Traditional approach vs Holistic Approach

- End to End Autonomous Driving, unlike Traditional Modular Stack, learns a single integrated policy.
- Most of these approaches emphasize on Planning Decisions to be informed directly by Raw Data.
- But the current E2E approaches lack interpretability and safety guarantees.
- But this is an attempt to break-down the common approaches followed inside the E2E Autonomous Driving Approaches.



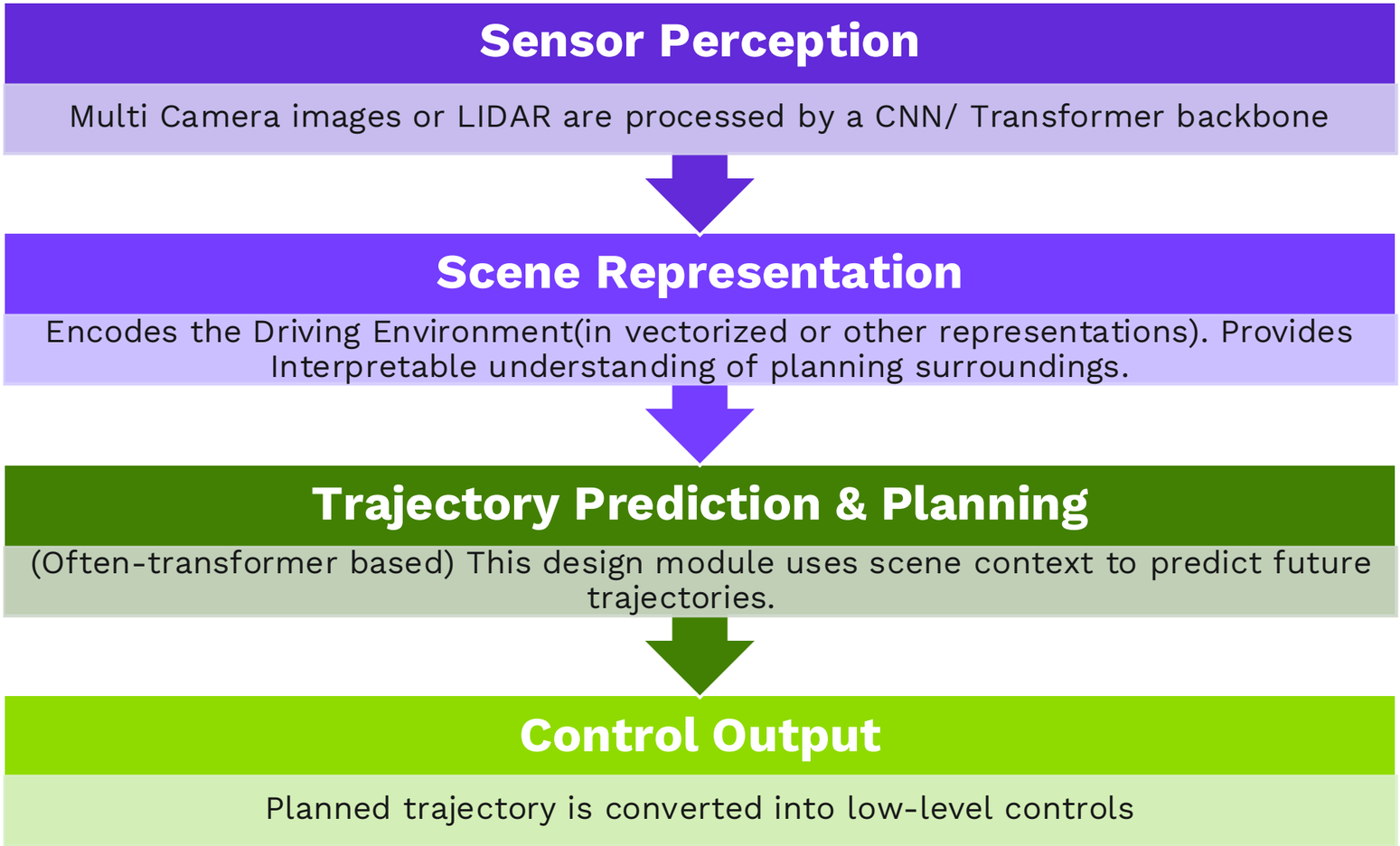
Universal flow of End-to-End Frameworks (super simplified)

History of E2E models (2019-2025)



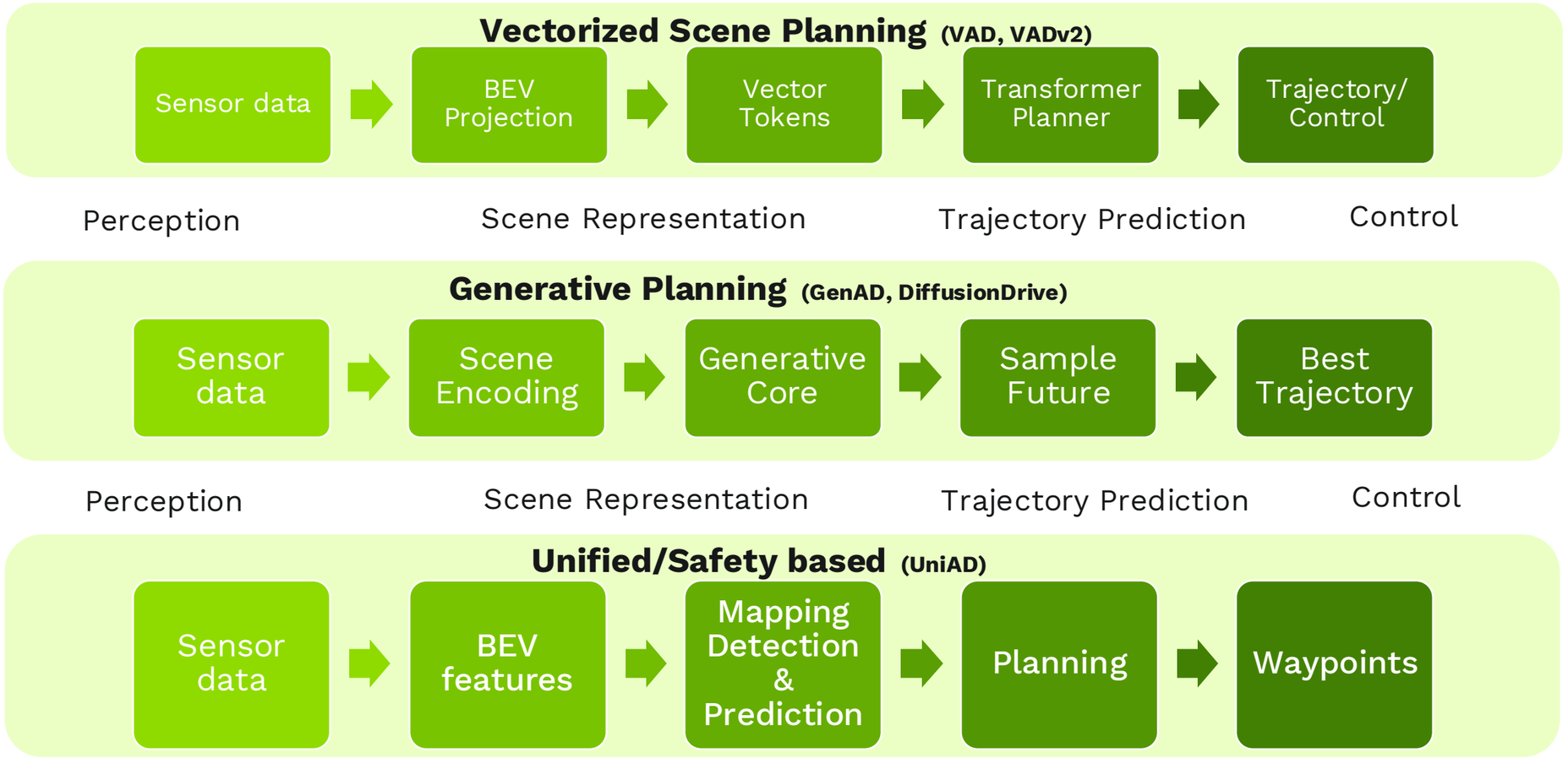
Modular structure of Not so modular Algorithms

What do the E2E stacks have in common?



Families in the Development

- The approaches in End-to-End Autonomous driving evolved into families with similar approaches alike. So far, the documentation includes transformer-based approaches.



Leading E2E Approaches and Best fit scenarios

Approach	Core Strength	Best Fit
Vectorized Planning	Plans Fast using lightweight lane & vehicle vectors	Cost-sensitive roll outs, edge devices
Generative Approaches	Imagines multiple futures before choosing a trajectory	Complex scenarios, Urban traffic
Unified/ Safety based	Blends human driving know-how with rule-based safety checks	Regulated markets, premium safety brands

Common Components & Key Differences

What is common amongst approaches and what's different?

1. **Scene Representation** : Interpretation is done by an E2E model using vectors/queries to perform structured reasoning. VAD uses vectors, GenAD uses latent representations, Diffusion drive uses truncated anchors.
2. **Modality in Planning:** Unlike traditional planners E2E stacks handle multi modal decisions like, having a score for "go left" vs "go straight". VADv2, Diffusion Drive take probabilistic route. GenAD, Diffusion Drive provide diverse trajectory options. UniAD, Hydra MDP focus on single "best" plan with extension queries and multiple futures.
3. **Training Strategy:** All these models learn from Data usually large driving sets. Hydra MDP blends imitation learning with RL, GenAD uses an aggregated 2000+ hours of video data.
4. **Family based differences :** VAD family emphasizes on Structured scene understanding and fast planning, GenAD, Diffusion Drive emphasizes on Future Prediction and capturing uncertainty. Unified and Multitask families emphasize incorporating domain knowledge.

Indian Datasets vs nuScenes

Dataset	Details	Inputs available	Annotations/labels available	Additional data
nuScenes	- 1000 20s scenes, - 1.4M camera images, - 390k LIDAR sweeps, - 1.4M RADAR sweeps - 1.4M object bounding boxes in 40k keyframes - Boston and Singapore	- RGB Images from 6 cameras, - PCD from 1 LIDAR, - PCD from 5 RADAR, - GPS, - IMU	- Semantic category, 3D bounding box and attributes for each object having atleast 1 lidar/radar point - Rasterized Semantic maps (road and sidewalks) - Baseline routes (without obstacles) - Ego_pose derived from LIDAR data	- Map expansion expands rasterized semantic maps with 11 classes through vectorized maps - Lidar semantic segmentation added in nuScenes-lidarseg - nuScenes and nuScenes-lidarseg combined in nuScenes-panoptic - CANbus messages (route, IMU, ego-pose)
IDD Segmentation	- 182 scenes/sequences, - 10000 images - Bengaluru and Hyderabad	- RGB Images from a stereo pair	- Image semantic segmentation into 34 classes spread across 4 levels of label hierarchy	- Not applicable
IDD-3D	- 12k annotated LIDAR frames - Hyderabad	- RGB images from 6 cameras - PCD from 1 LIDAR	- LIDAR frames annotated with 3D bounding boxes around objects labelled with 10 main and 7 extra classes	- Not applicable
IDD Detection	- 40000 images		- Bounding box annotations	- Not applicable

Indian Datasets vs nuScenes (contd.)

Dataset	Details	Inputs available	Annotations/labels available	Additional data
DriveIndia	- 66986 RGB images - Hyderabad	- RGB Images from multiple cameras	- 2D bounding boxes for objects with labels from 24 classes	- Not applicable
IDD Multi-modal	- Stereo images from front camera (15FPS) - GPS Lat long (15Hz) - LIDAR and OBD	- NA	- NA	- NA

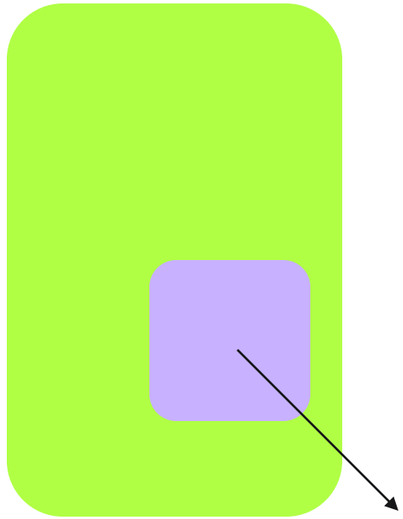
nuScenes for E2E AD Stacks

Stack	Inputs used	Annotations/labels used
UniAD	Multi-camera Images **only**	- Perception: • Class labels, 3D bounding boxes and object IDs for **Detection** and **Tracking**, • Semantic maps (from Map extension) for **Online Mapping** - Prediction: • *3D bounding box positions converted to trajectories and object speeds for smoothening* * for **Motion forecasting** • *3D bounding boxes (mapped to BEV)** for **Occupancy prediction** - Planning: *Ego-pose and 3D bounding boxes of surrounding objects** (to avoid collisions) for **Planning**
GenAD	TBD	TBD
VAD	TBD	TBD

*: derived information from the paper and the repo; not explicitly stated

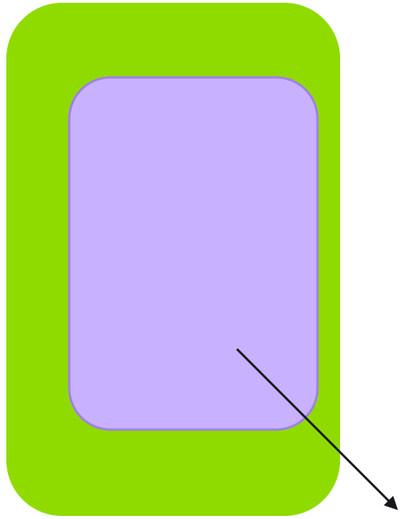
nuScenes for E2E AD Stacks (contd.)

NuScenes Inputs (images, LIDAR and RADAR points, CAN data, ego pose)



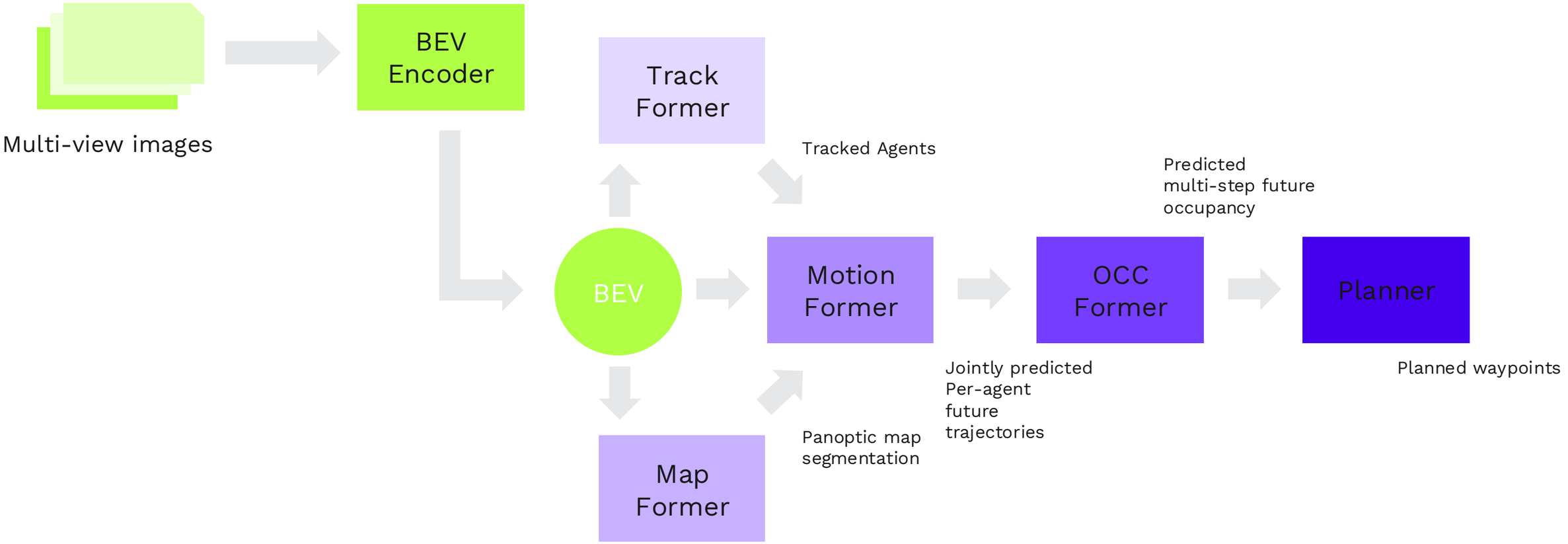
UniAD uses a subset of the inputs available in the nuScenes dataset

NuScenes annotations (3D bounding boxes, classes, attributes, semantic maps, ego pose)



UniAD uses a subset of the annotations provided by the nuScenes dataset, larger in proportion to the inputs used

UniAD at a glance



UniAD: Model complexity and computational costs*

Tasks	#Params	FLOPs	FPS
Detection (BEVFormer), tracking, mapping,motion prediction, occupancy prediction and planning	125.0M	1.709T	1.8

Image backbone	#Params	FLOPs	FPS
ResNet50	63.46M	0.025T	25
ResNet101	82.45M	0.033T	17.5
VoVNetV2-99	TBD	TBD	TBD

Memory Requirements:

- Inference -> for a 1B parameter model, memory requirements
  - 32 bit float – 4GB
  - 16 bit float – 2GB
  - Int8 precision – 1GB
- Training -> 4X Inference**

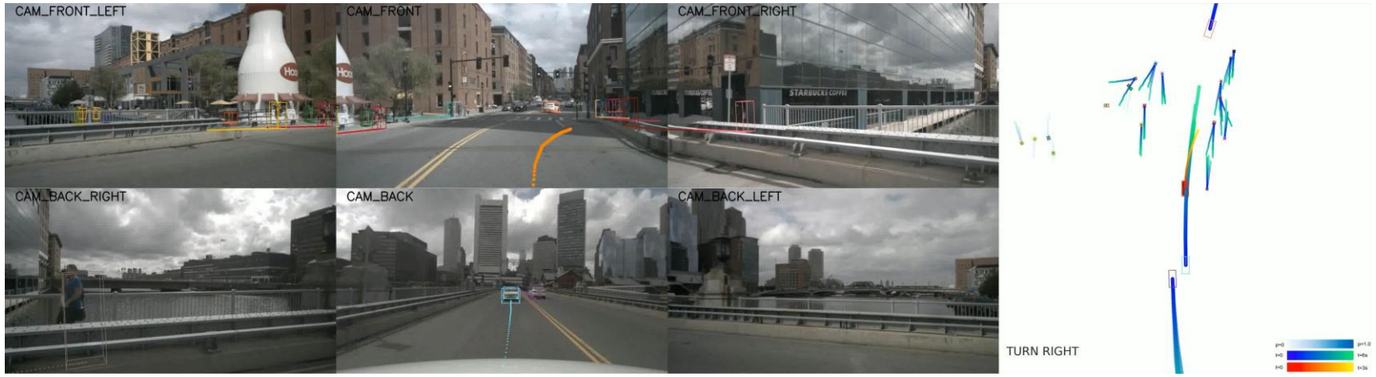
*on Nvidia Tesla A100
**https://medium.com/@plthiyagu/llm-model-parameter-memory-required-for-training-and-inference-634963b36b59

UniAD: Model demo on nuScenes v1.0_mini

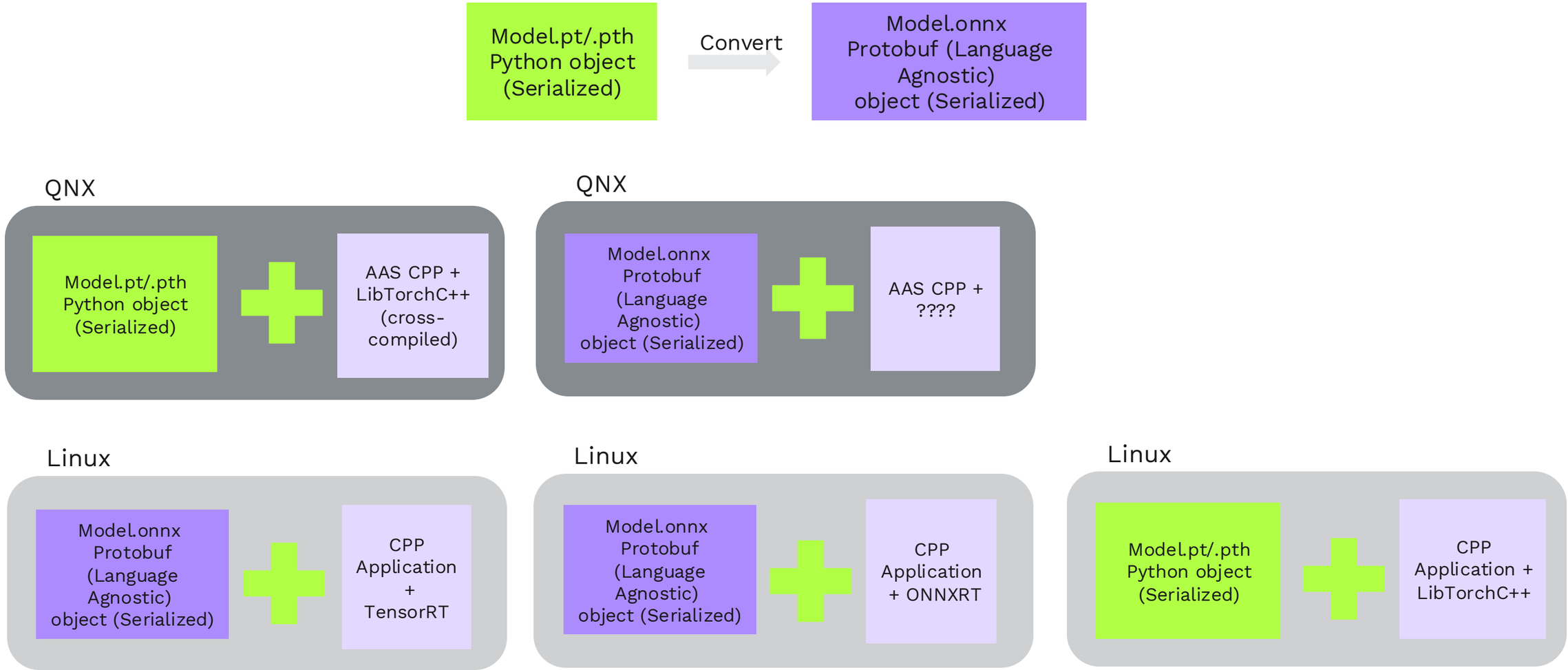
v1.0_mini

- A mini version of the dataset with 10 scenes, 8 from training split and 2 from validation split of the larger dataset
- 6 scenes from Singapore and 4 scenes from Boston

Boston-Seaport and Singapore-Queenstown



Porting to Production HW



References

[1] VAD: Vectorized Scene Representation for Efficient Autonomous Driving

[2] VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning

[3] GenAD: Generative End-to-End Autonomous Driving

[4] Hydra-MDP: End-to-end Multimodal Planning with Multi-target Hydra-Distillation

[5] Scene-Adaptive Motion Planning with Explicit Mixture of Experts and Interaction-Oriented Optimization

[6] DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

[7] nuScenes: A multimodal dataset for autonomous driving