NVIDIA

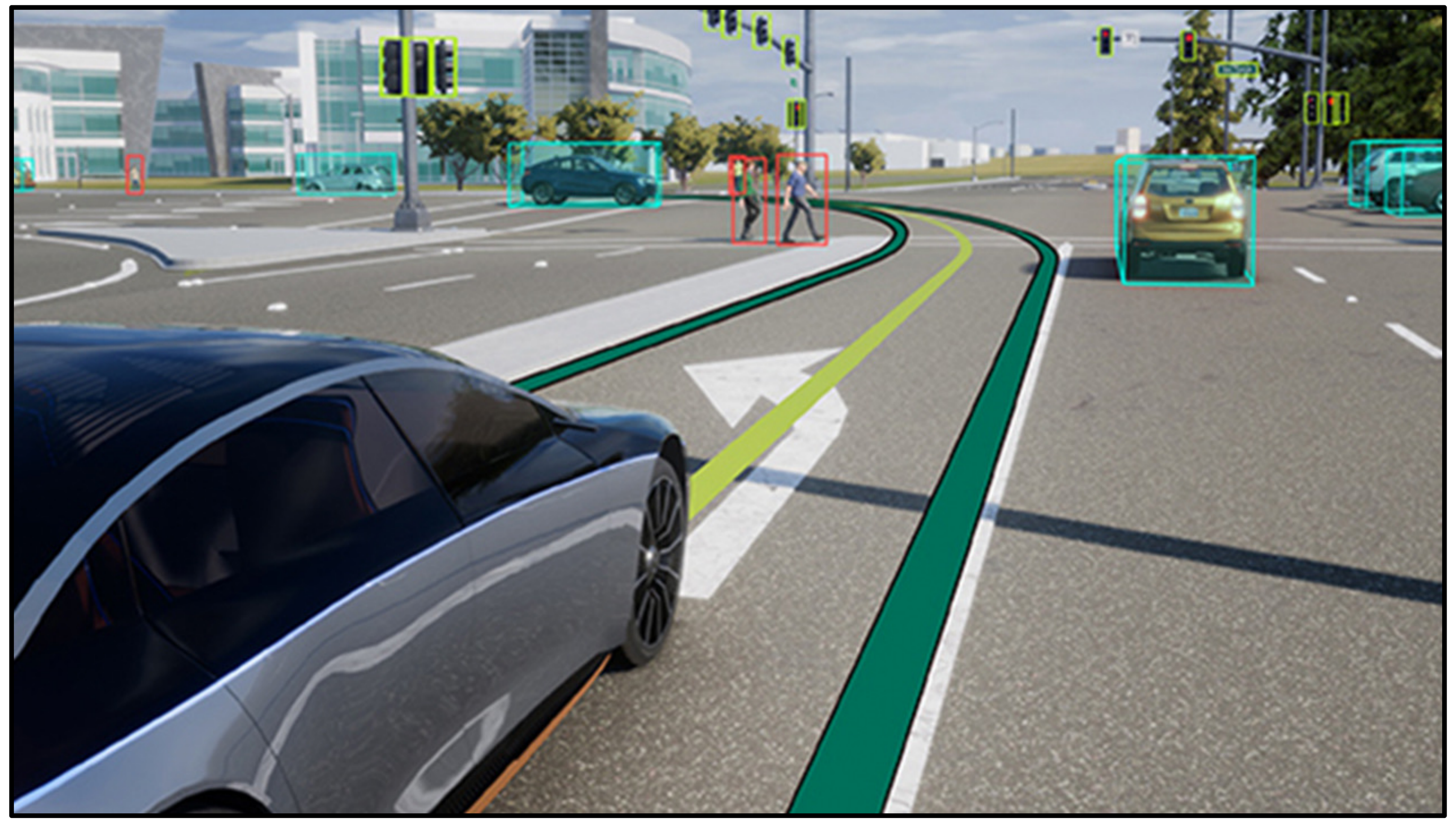
Revolutionizing AV Development with Foundation Models

Boris Ivanovic | NVIDIA Autonomous Vehicles Research Group

AVIATE Seminar | April 19th, 2024

Where is the AV industry today?

- Advanced driver assistance systems (ADAS) in the hands of consumers
- Driverless operations in certain operational design domains (ODDs)



Powered by “judicious incorporation” of ML/AI into the AV stack:

Safety is paramount

What do we need for tomorrow?

New deployment domains…

Optical illusions…

…???

There’s still a “long tail” of anything/everything that could possibly happen out there!

How can we *generalize* to the unseen?

How do Humans Navigate the “Long Tail”?

By leveraging strong contextual understanding and common-sense reasoning

Contextual Understanding

Reasoning Ability

Good to Go!

<u>Generalist Experience</u>
Lifetime of interacting with the world beyond driving

+

<u>AV Fleet</u>
(Many) lifetimes of driving data: costly to acquire, yet still incomplete

=

A next generation of AVs equipped to handle new domains + the long tail?

How can we access this generalist experience for AV applications?

Providing AVs with Lifetimes of *Generalist* Experience

AV Foundation Model

Long-Tail Generalization via Common-Sense Reasoning

Offline Processes

Autolabeling
Sensor and Traffic Simulation
Scenario Generation
…

Safety Frameworks

On-Vehicle AV Stack

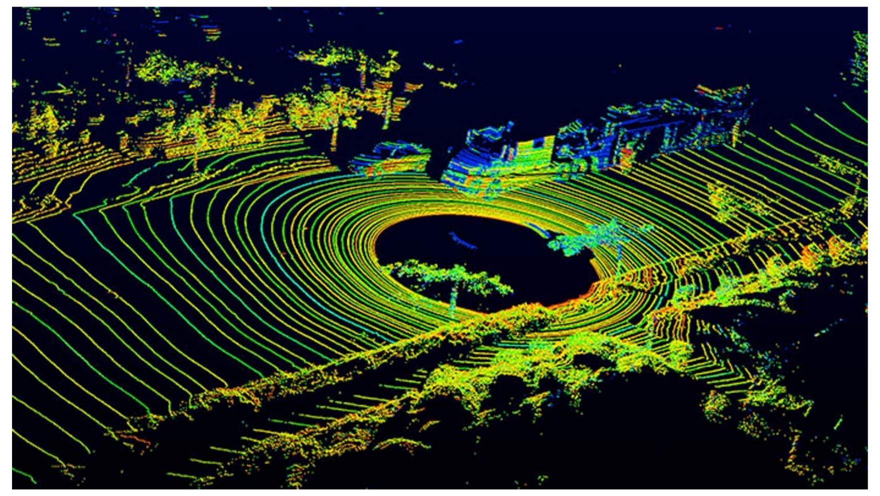
FM-empowered E2E Stacks
In-Cabin Driving Assistance
Interactive Vehicle Interface
…

Building AV Foundation Models

Multimodal Foundation Models

Learning universal representations in a shared embedding space

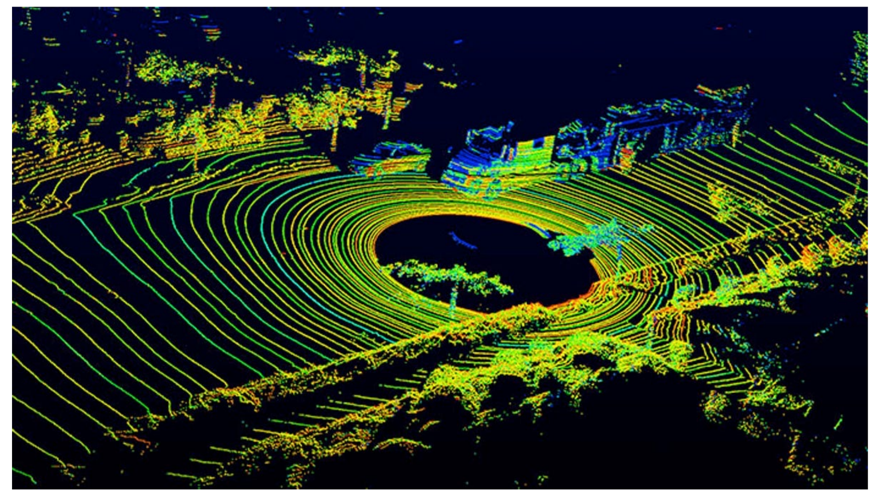
COMMON CRAWL



Multimodal Foundation Models

Learning universal representations in a shared embedding space

COMMON CRAWL



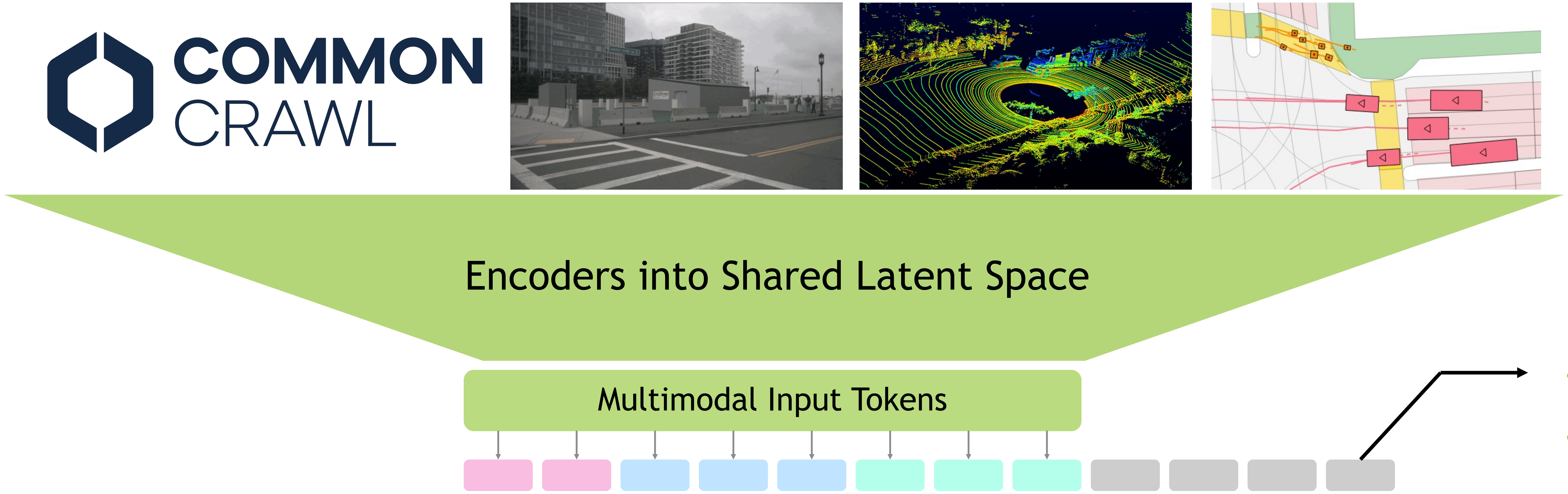
Encoders into Shared Latent Space

What is an "encoder"?

- Neural network that extracts features from raw data

Multimodal Foundation Models

Learning universal representations in a shared embedding space



What is a "token"?

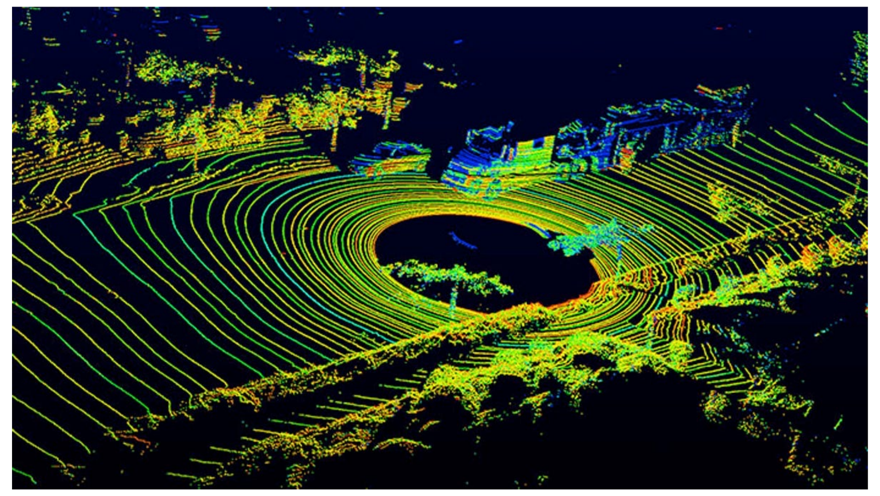
- Semantic unit of language,
- Image patch embedding,
- Video/lidar/radar/etc. sensor embedding,
- Vehicle state, action, trajectory embedding,
- Latent scene representation
- … anything!

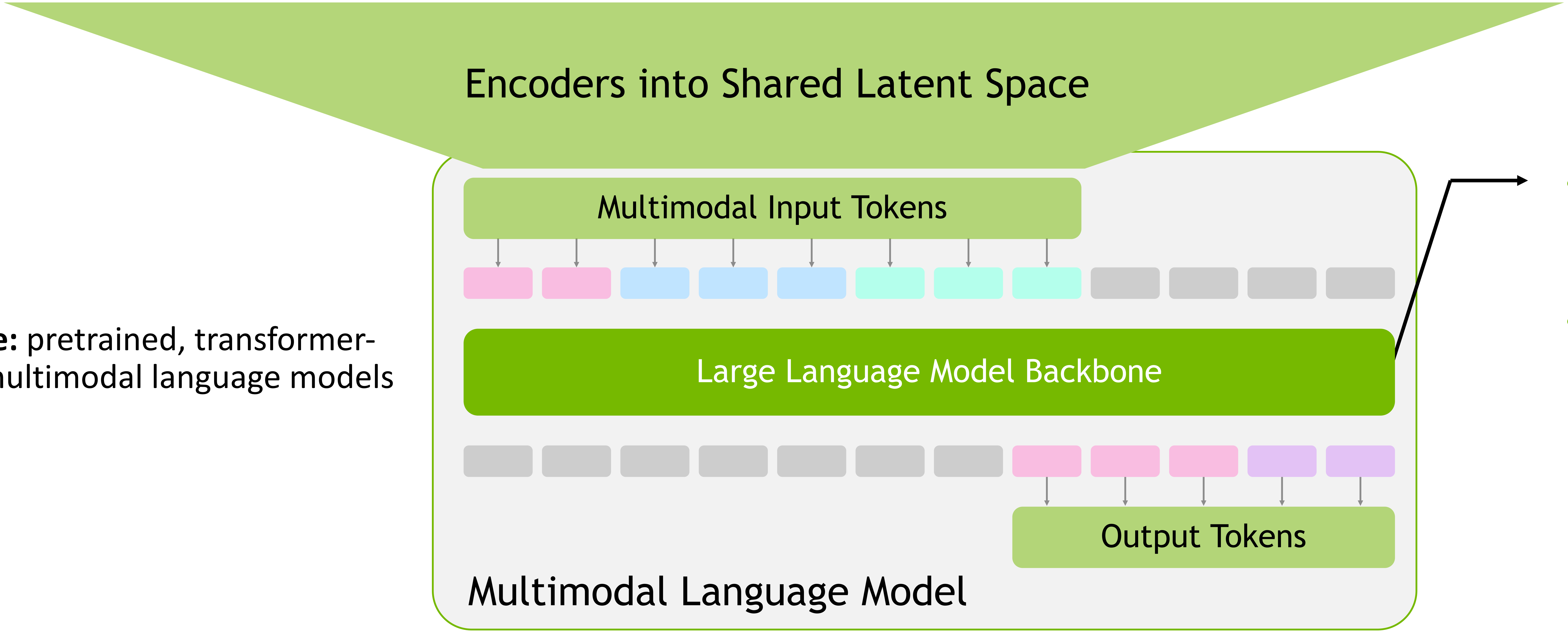
"Unit of information" from arbitrary modality

Multimodal Foundation Models

Learning universal representations in a shared embedding space

COMMON CRAWL





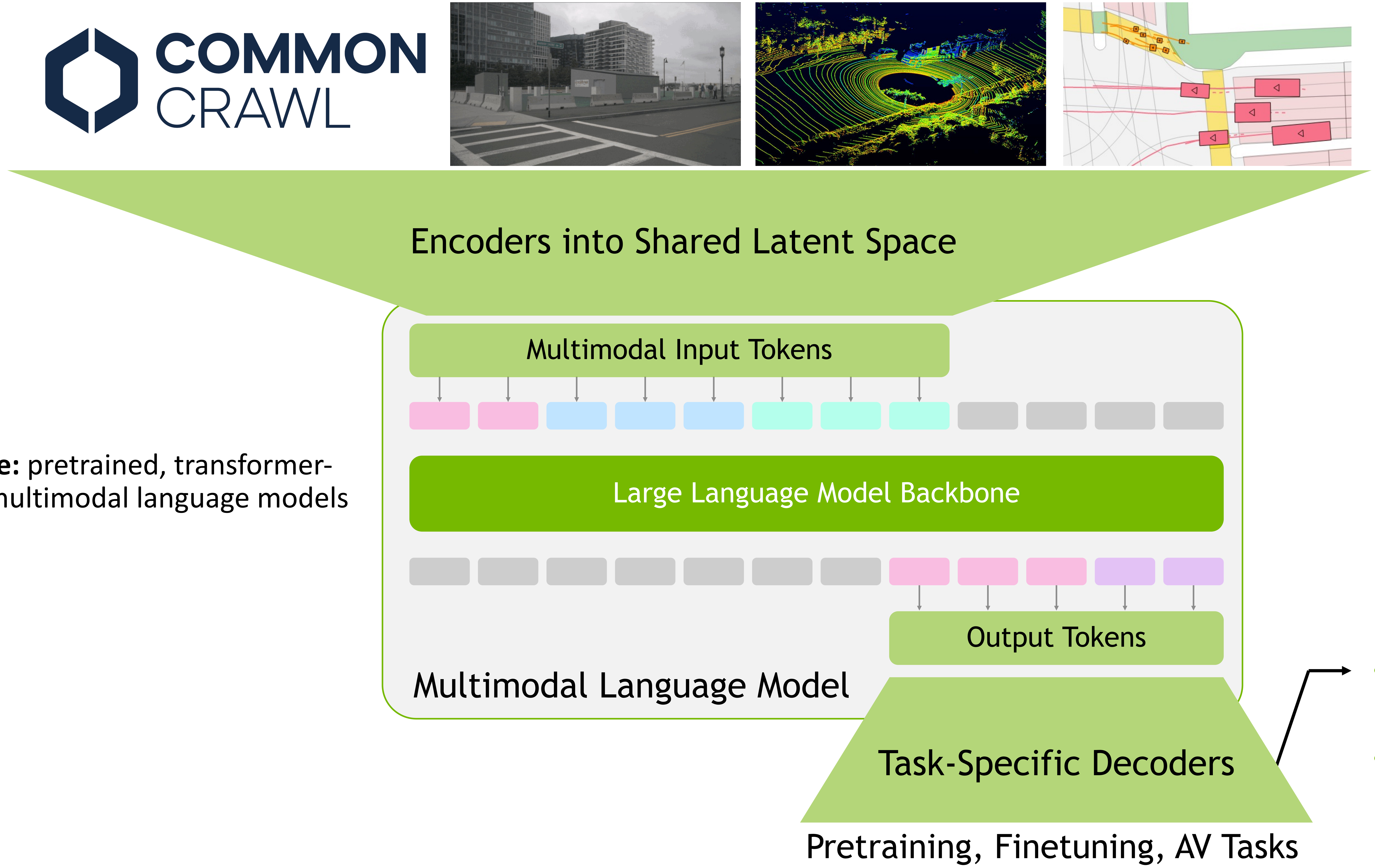
Example: pretrained, transformer-based multimodal language models

What is a “backbone”?

- Neural network that processes input tokens and generates output tokens.
- In this example, an existing LLM is used to process tokens autoregressively.

Multimodal Foundation Models

Learning universal representations in a shared embedding space



Example: pretrained, transformer-based multimodal language models

What is a “decoder”?

- Converts output tokens to modalities of interest.
- E.g., bounding boxes, trajectories, maps, images, videos, etc.

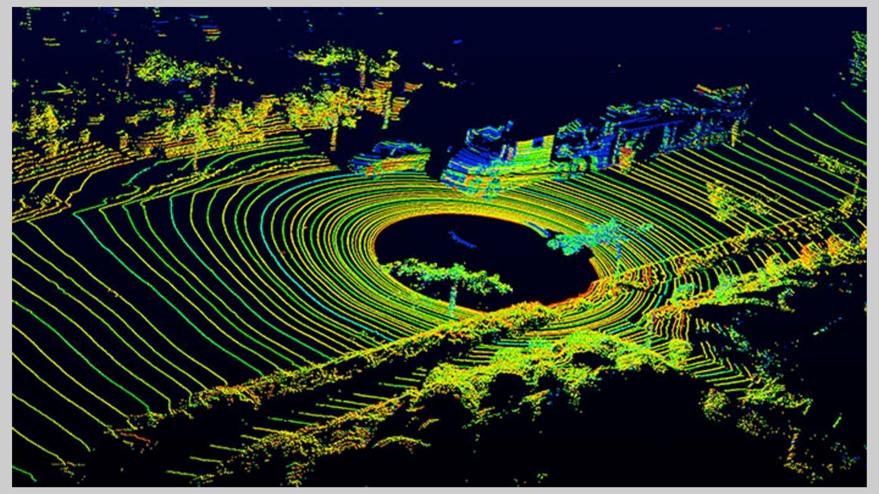
How do we Build an AV FM?

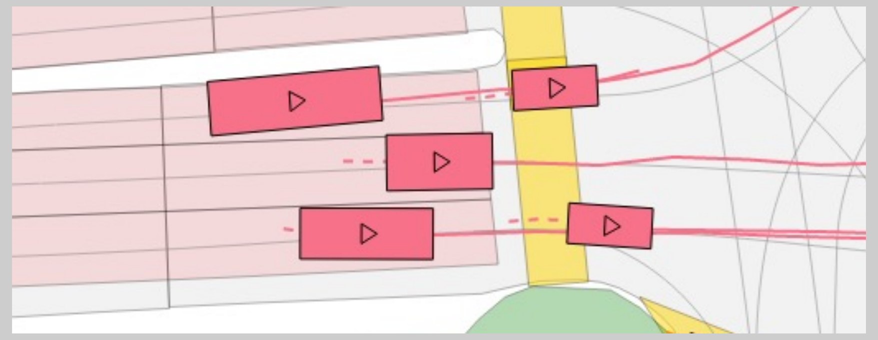
Desired capabilities inform choice of data, model, and training tasks

Data

COMMON CRAWL

Internet-scale Data





AV-specific data

+

Model

Representation

- Text Tokens
- Discrete Visual Tokens
- Ego Perspective Features
- 3D Features

Backbone

- Autoregressive Transformer
- Diffusion Model
- Hybrid

+

Self-supervised Task (Pretraining)

- Text Completion
- Image / Video Generation
- Depth Estimation
- Scene Flow Estimation
- Image / Video Captioning

Supervised Task (Finetuning)

- OCR
- Grounding
- VQA (MMMU)
- Detection
- Tracking
- Planning
- Common Sense Reasoning

Multi-Modal Large Language Models

Li, Wang, Mao, Ivanovic, Veer, Leung, Pavone, *Driving Everywhere with Large Language Model Policy Adaptation*, CVPR 2024

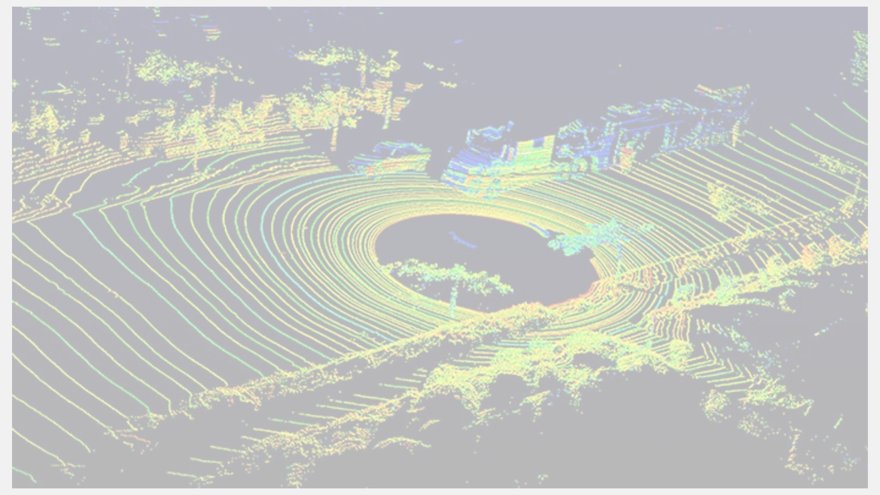
Cho, Ivanovic, Cao, Wang, Schmerling, Weng, Li, You, Kraehenbuehl, Wang, Pavone, *Language-Image Models with 3D Understanding* (submitted)

Video Generation Models

Data

COMMON CRAWL

Internet-scale Data



AV-specific data

Model

Representation

- Text Tokens
- Discrete Visual Tokens
- Ego Perspective Features
- 3D Features

Backbone

- Autoregressive Transformer
- Diffusion Model
- Hybrid

Sora

DALL·E 3

STABLE DIFFUSION

Unsupervised Task (Pretraining)

- Text Completion
- Image / Video Generation
- Depth Estimation
- Scene Flow Estimation
- Image / Video Captioning

Supervised Task (Finetuning)

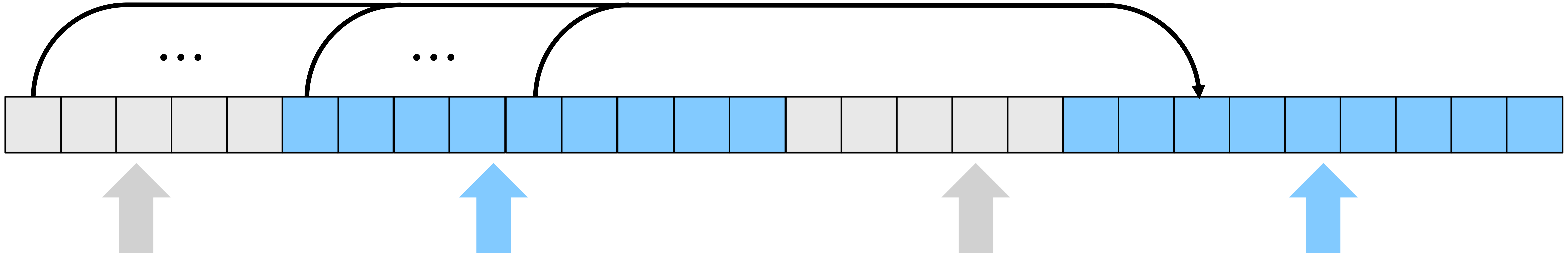
- OCR
- Grounding
- VQA (MMMU)
- Detection
- Tracking
- Planning
- Common Sense Reasoning

Case Study: Video Generation via Tokens

Architecture and Potential Training Tasks

1. General image and video generation on internet data
2. Multi-camera video generation with AV data
3. Traffic simulation, simulating bounding box trajectories conditioned on ego actions
4. Sensor simulation, simulating camera images and LiDAR returns from bounding boxes

Each can be additionally conditioned on associated text prompts/captions and sensor calibration info



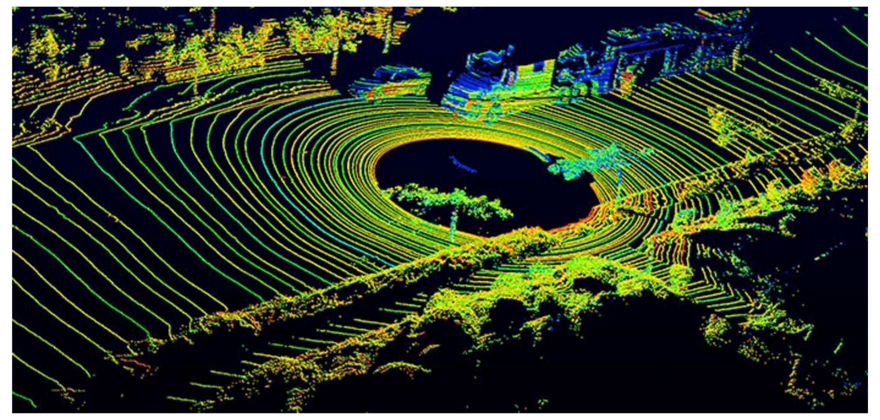
• • •

Embed (optional) conditioning information

- Sensor calibration
- Text description
- Ego vehicle next action
- 2D/3D bounding boxes

…

Tokenize raw sensors

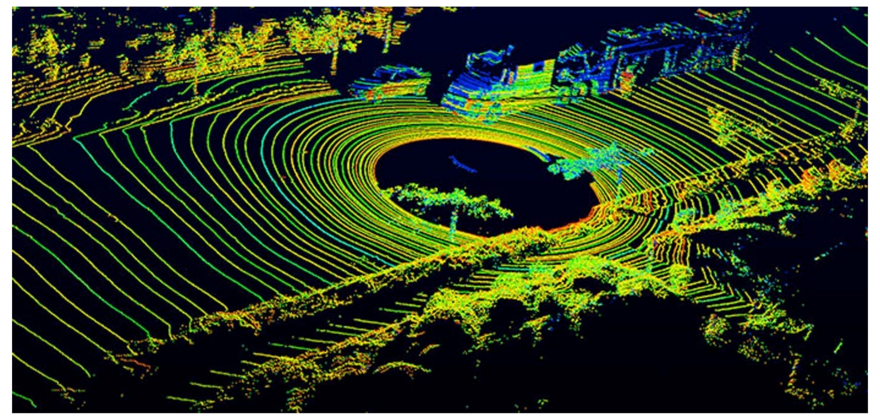


Embed (optional) conditioning information

- Sensor calibration
- Text description
- Ego vehicle next action
- 2D/3D bounding boxes

…

Tokenize raw sensors



• • •

Cross-Functional VFM-AV Team led by Jonah Philion

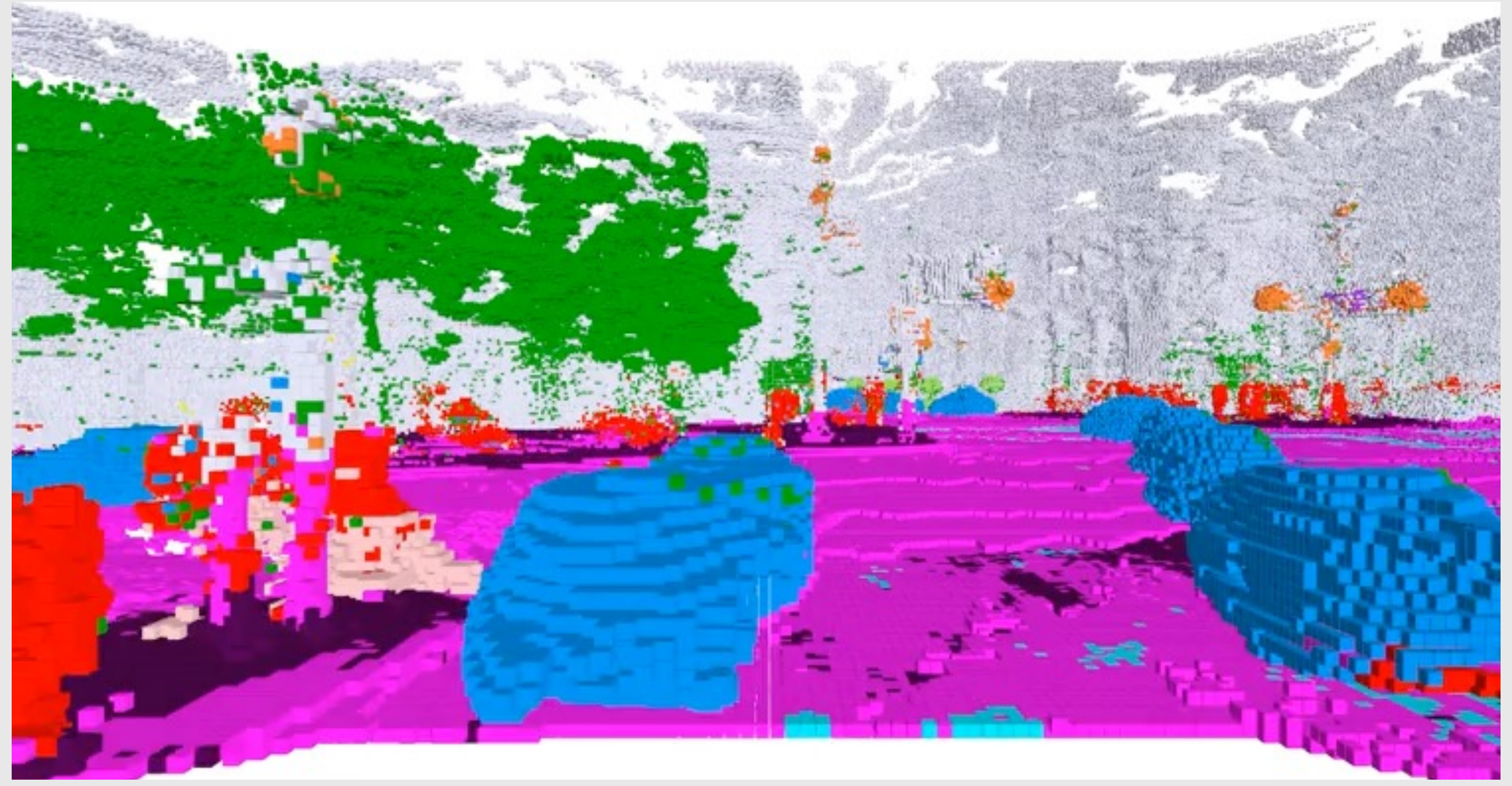
Case Study: Multi-Camera Video Generation via Tokens

Cross-Functional VFM-AV Team led by Jonah Philion

Using AV Foundation Models

How Can We Use AV FMs?

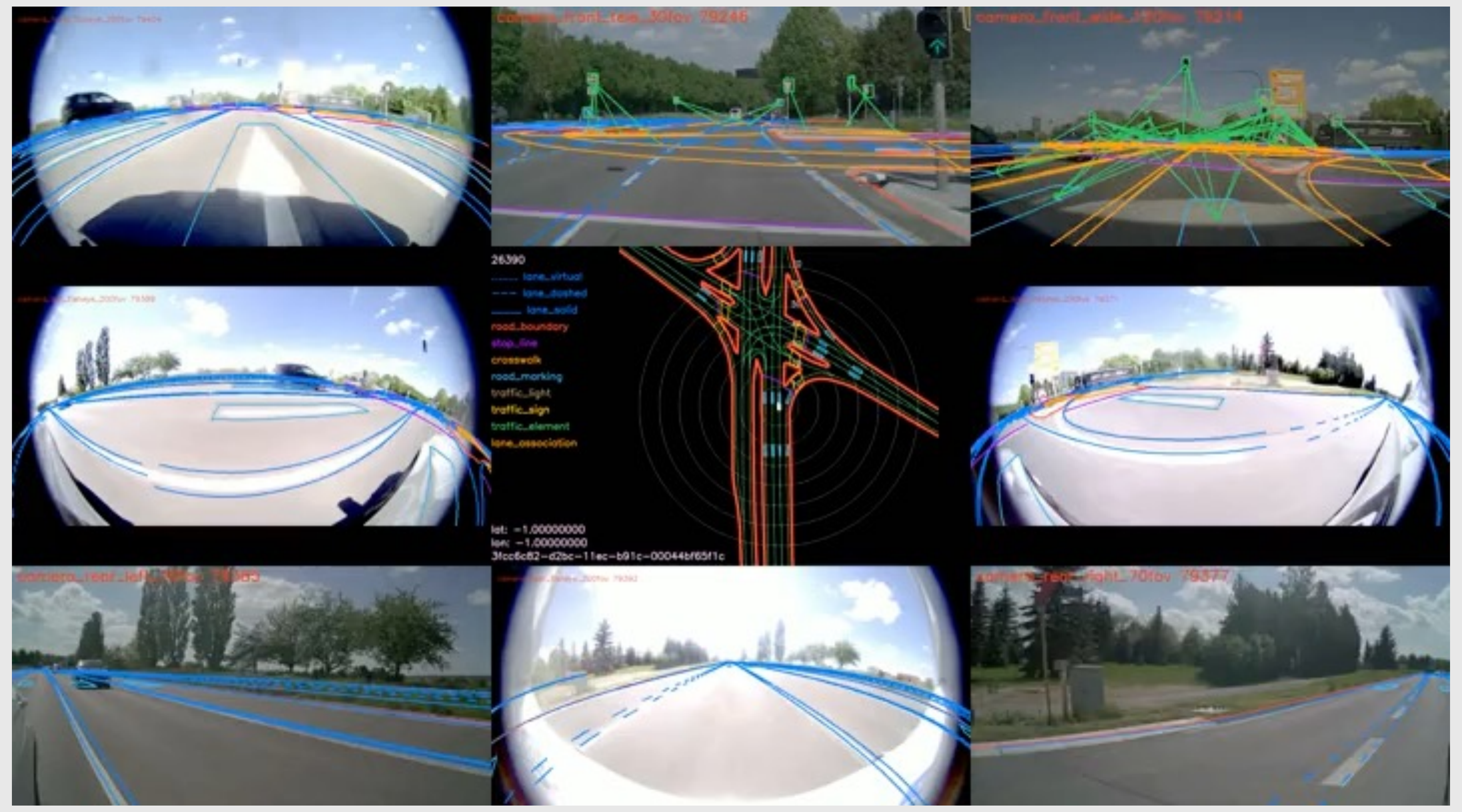
Offline Processes



Autolabeling

Simulation

On-Vehicle AV Stack

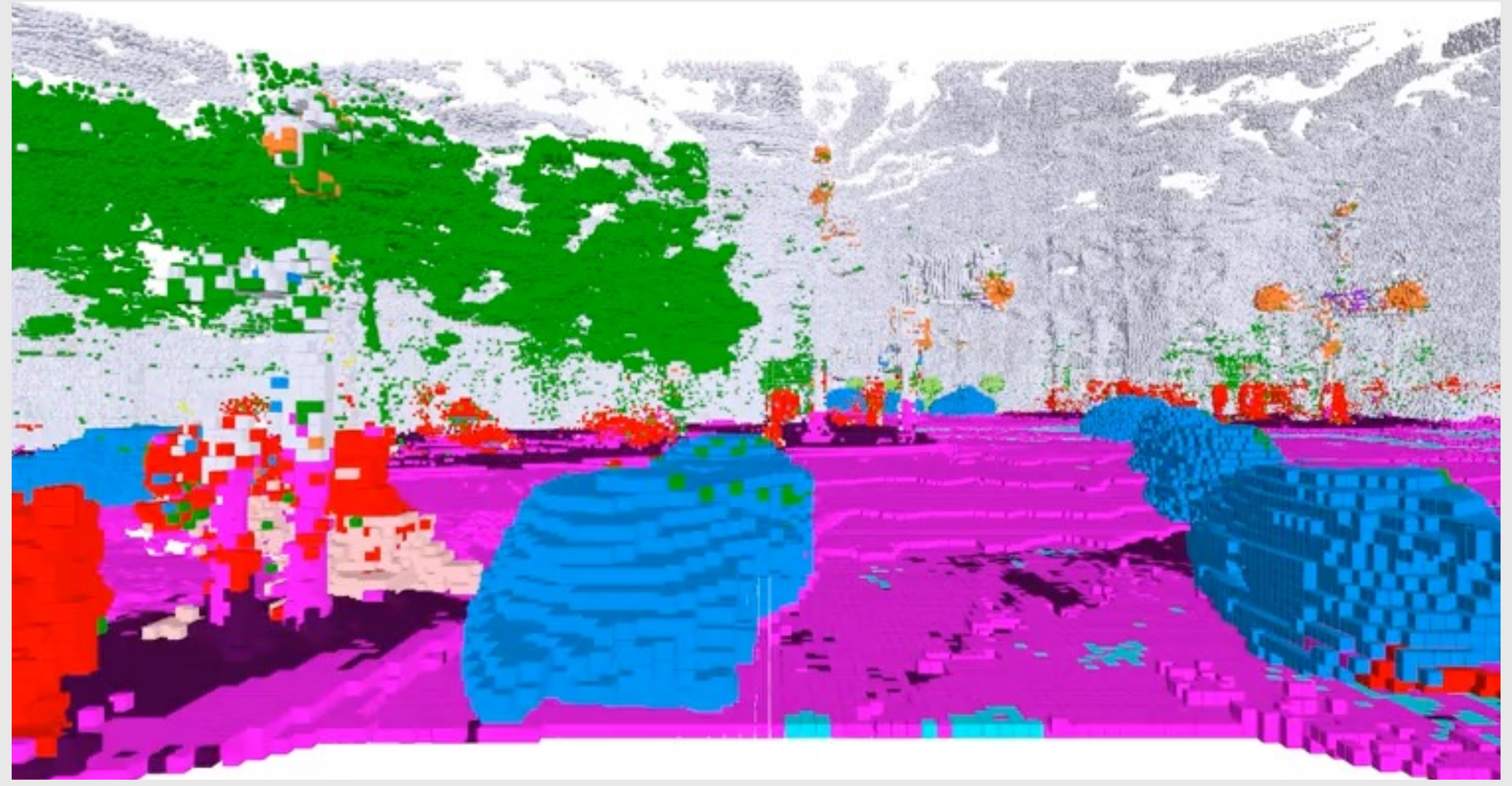


Novel End-to-End Architectures

In-Cabin Assistance

How Can We Use AV FMs?

Offline Processes



Autolabeling

Simulation

On-Vehicle AV Stack

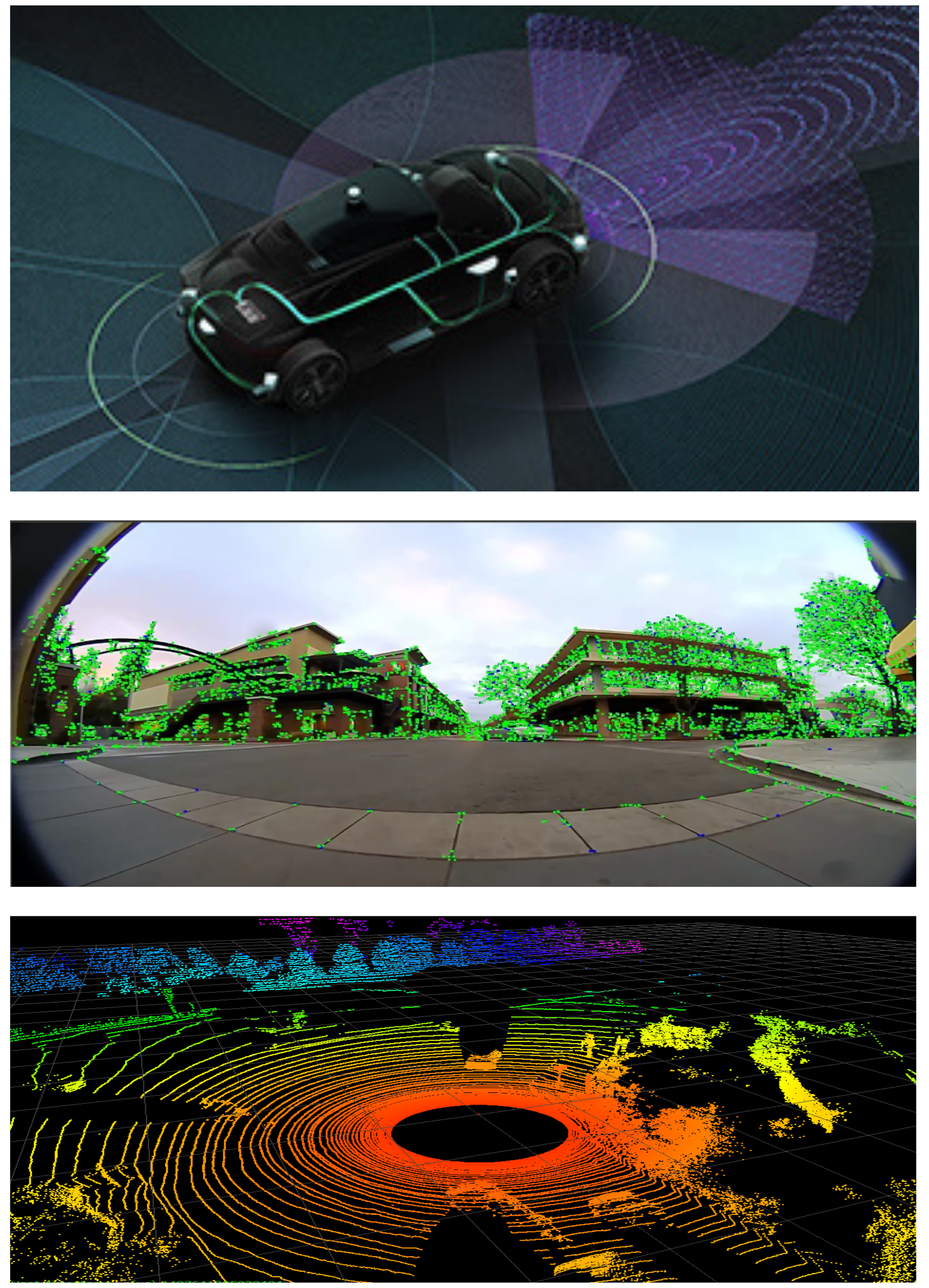
Novel End-to-End Architectures

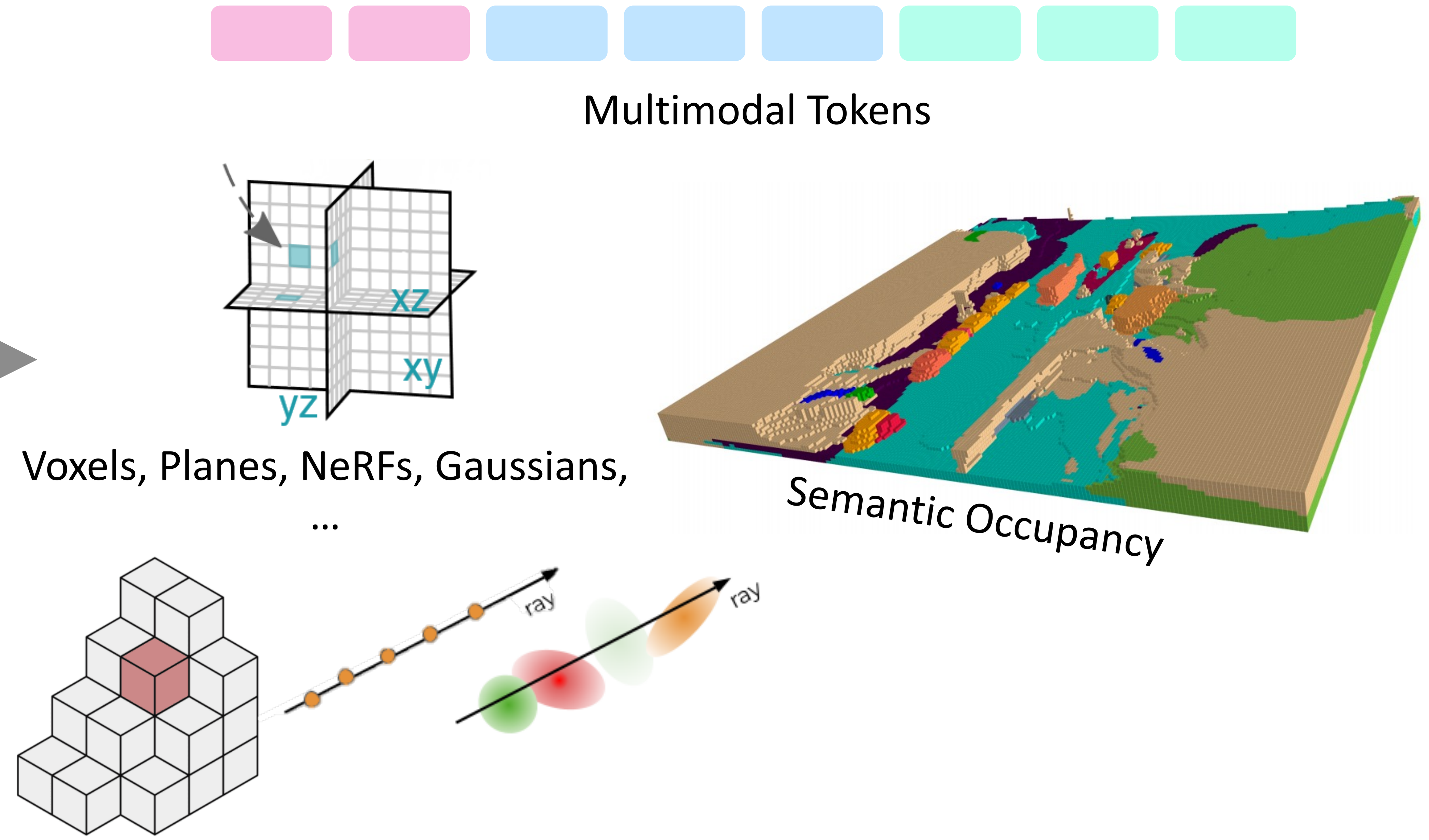
In-Cabin Assistance

Imbuing World Representations with Internet-Scale Priors

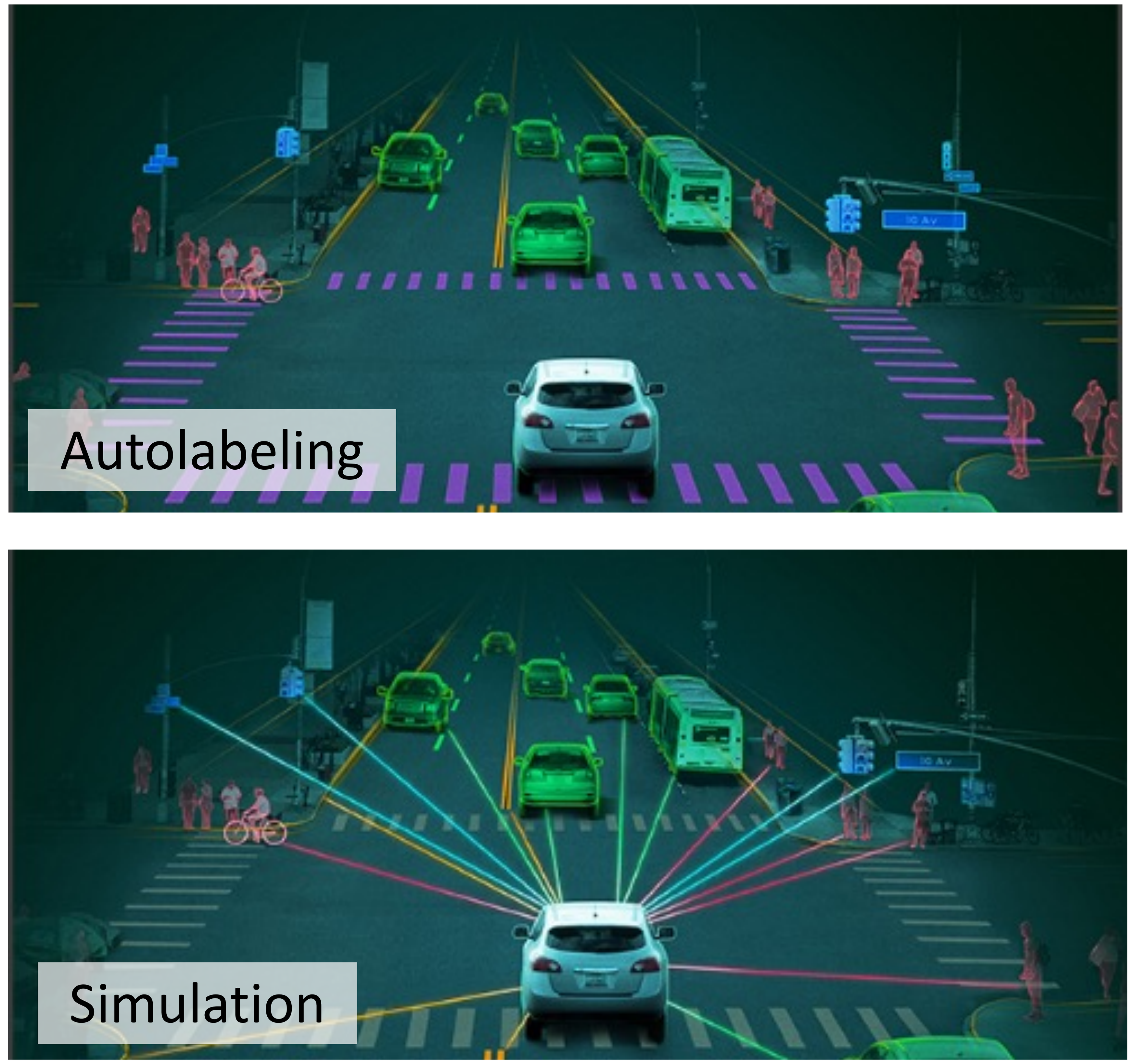
AV Foundation Model

General Semantic Features
Common-Sense Reasoning





World Representations



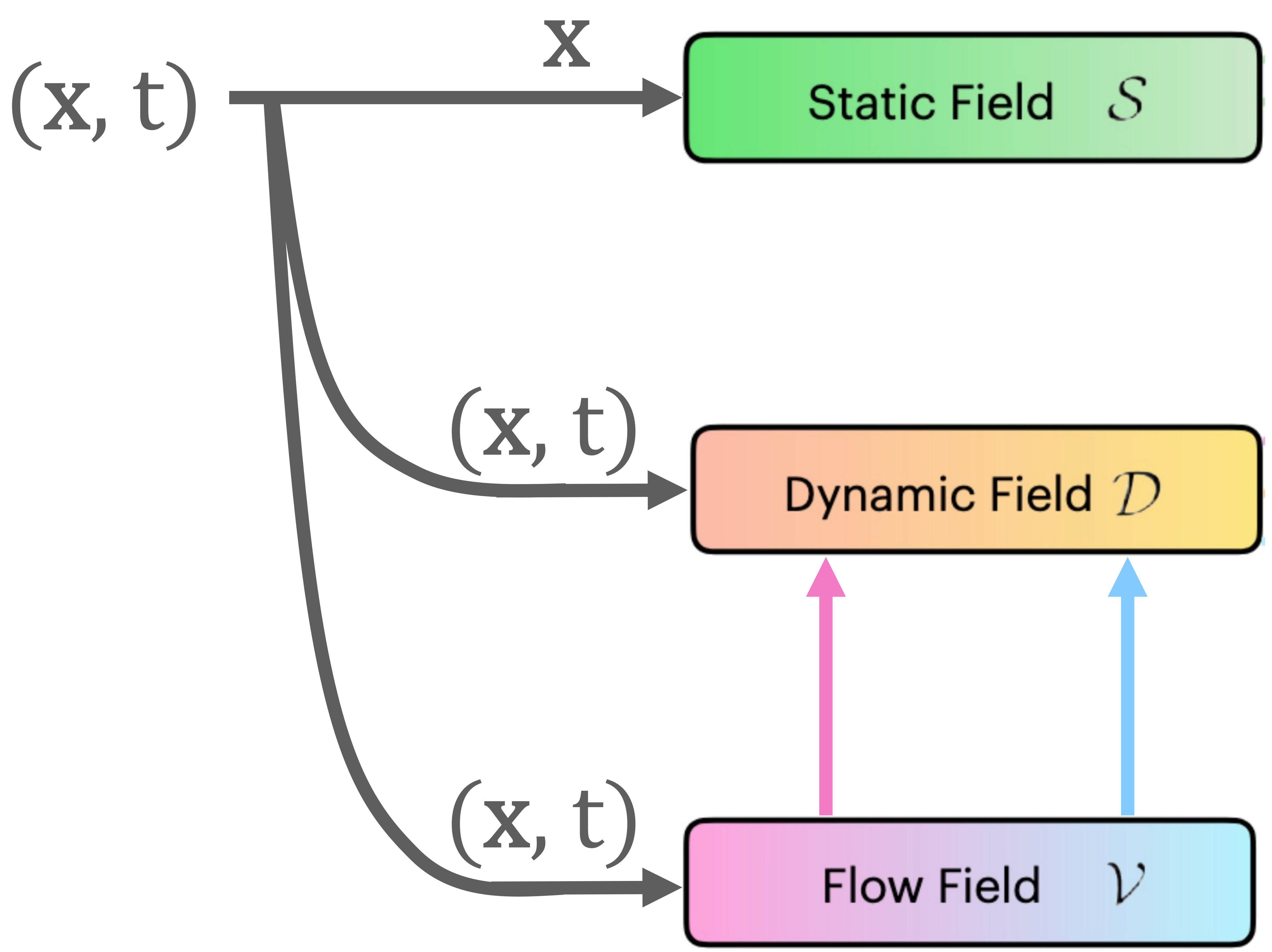
Dynamic Driving Scene Reconstruction and Representation

Static-Dynamic Decomposition

Yang, Ivanovic, Litany, Weng, Kim, Li, Che, Xu, Fidler, Pavone, Wang, *EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision*, ICLR 2024 https://emernerf.github.io/

Dynamic Driving Scene Reconstruction and Representation

Static-Dynamic Decomposition



Ground Truth Cameras

Yang, Ivanovic, Litany, Weng, Kim, Li, Che, Xu, Fidler, Pavone, Wang, *EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision*, ICLR 2024

https://emernerf.github.io/ Drive Labs: youtube.com/watch?v=4Ort_bdTQlk

General Semantic Representations with Foundation Model Features

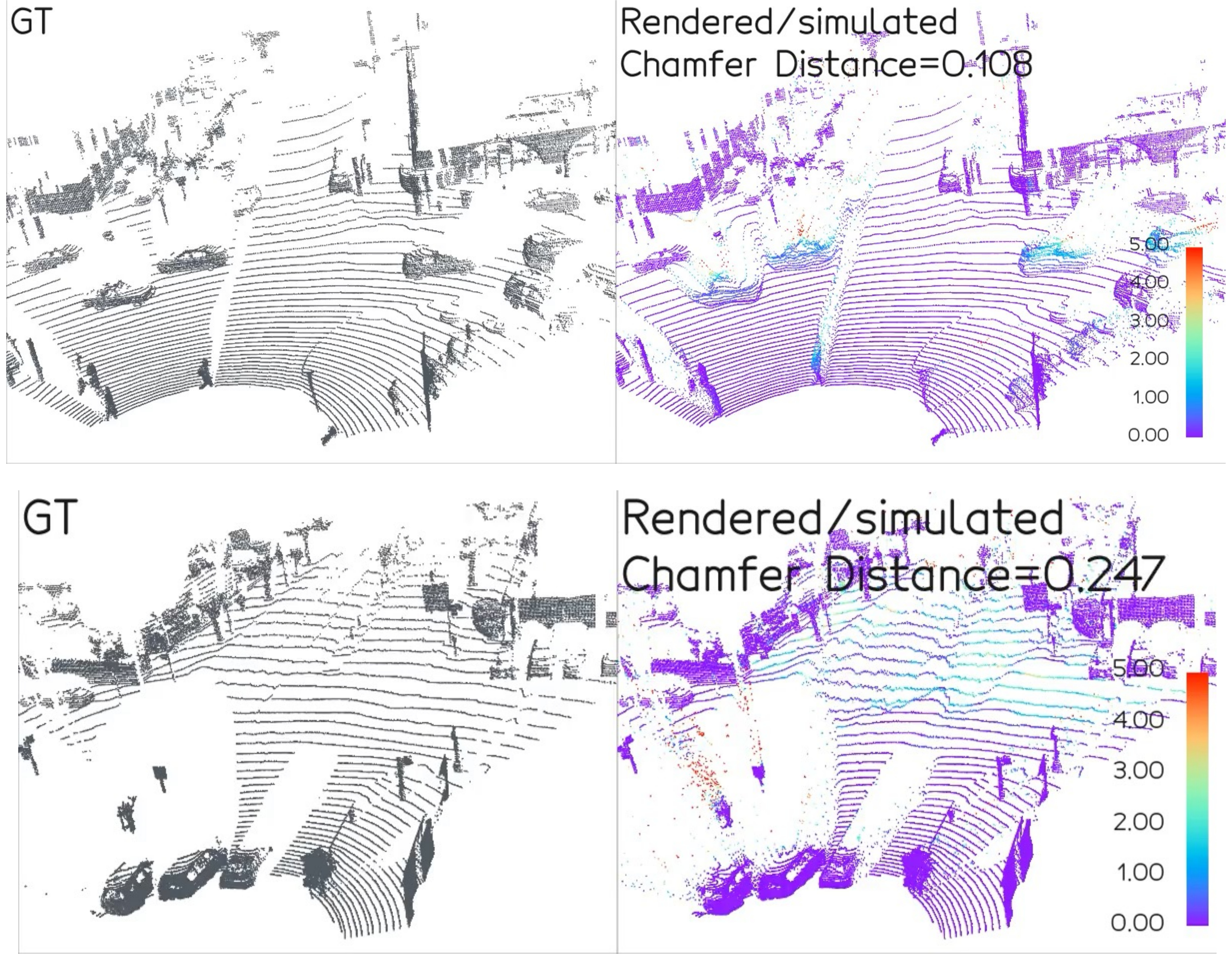
Autolabeling

Yang, Ivanovic, Litany, Weng, Kim, Li, Che, Xu, Fidler, Pavone, Wang, *EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision*, ICLR 2024 https://emernerf.github.io/

Neural Rendering for High-Fidelity Sensor Simulation

Original Camera Log

Rendered Camera Log



LiDAR Simulation

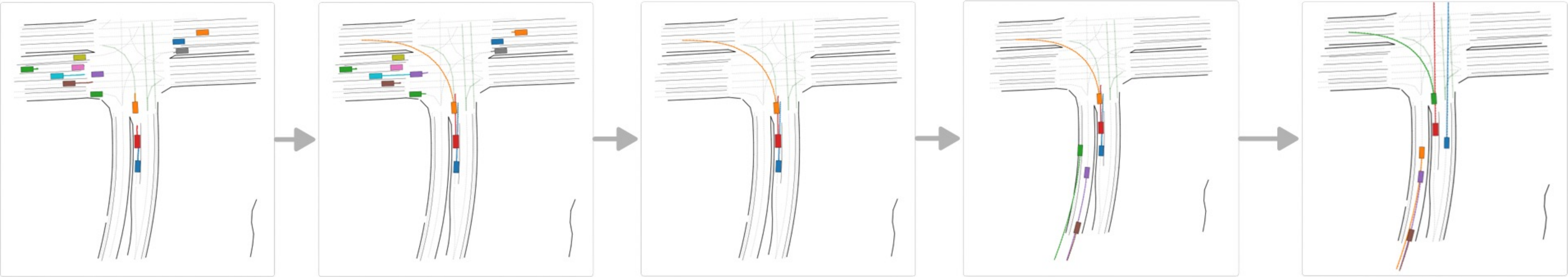
Yang, Ivanovic, Litany, Weng, Kim, Li, Che, Xu, Fidler, Pavone, Wang, *EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision*, ICLR 2024 https://emernerf.github.io/

Accelerating Traffic Simulation with Foundation Models

Xu, Chen, Ivanovic, Pavone, *BITS: Bi-level Imitation for Traffic Simulation*, ICRA 2023
Zhong, Rempe, Xu, Chen, Veer, Che, Ray, Pavone, *Guided Conditional Diffusion for Controllable Traffic Simulation*, ICRA 2023
Tan, Ivanovic, Weng, Pavone, Kraehenbuehl, *Language Conditioned Traffic Generation*, CoRL 2023
Zhong, Rempe, Chen, Ivanovic, Cao, Xu, Pavone, Ray, *Language-Guided Traffic Simulation via Scene-Level Diffusion*, CoRL 2023 (Oral)
Chen, Tonkens, Pavone, *Categorical Traffic Transformer: Interpretable and Diverse Behavior Prediction with Tokenized Latent*, arXiv 2023
Ding, Cao, Zhao, Xiao, Pavone, *RealGen: Retrieval Augmented Generation for Controllable Traffic Scenarios*, arXiv 2023

Accelerating Traffic Simulation with Foundation Models

Transforming text to simulation



Input | "make the car in front turn left" | "remove all the horizontal cars" | "add more cars on the left" | "speed up same-direction cars"

"Vehicle 1 did not notice that traffic was slowing down and struck the rear of Vehicle 2." *NHTSA CIREN #154*

"Vehicle 1 slowed down to turn left into a private driveway. Vehicle 2 tried to pass in the oncoming lane, striking Vehicle 1." *NHTSA CIREN #806*

"As Vehicle 1 was passing through the intersection, Vehicle 2 turned left, striking Vehicle 1." *NHTSA CIREN #324*

Tan, Ivanovic, Weng, Pavone, Krähenbühl, *Language Conditioned Traffic Generation*, CoRL 2023 https://ariostgx.github.io/lctgen/

Accelerating Traffic Simulation with Foundation Models

Transforming text to simulation

As Vehicle 1 approached the intersection, its driver did not notice the vehicles stopped ahead at the traffic light. The traffic signal turned green and Vehicle 2 began to slowly move forward. The frontal plane of Vehicle 1 struck the rear plane of Vehicle 2 ...

→ **GPT4** →

Summary: V1 approaches an intersection and does not notice V2 ahead...
Explanation: [V1] - Because V1 is moving , we assume V1's initial speed is 10 m/s (index 4). V1 keeps going straight, so its actions are all 4 (keep speed). [V2] - As V1 is moving straight and hits V2 from behind, V2 is in front of V1....
Output:
- 'V1': [-1, 0, 0, 4, 4, 4, 4, 4] - 'V2': [3, 8, 2, 0, 4, 4, 4, 4] - 'Map': [2, 2, 2, 2, 8, 1]

Tan, Ivanovic, Weng, Pavone, Krähenbühl, *Language Conditioned Traffic Generation*, CoRL 2023 https://ariostgx.github.io/lctgen/

Simultaneous Sensor *and* Traffic Simulation

Camera Log of the Scenario

Rendered Camera in New Scenario

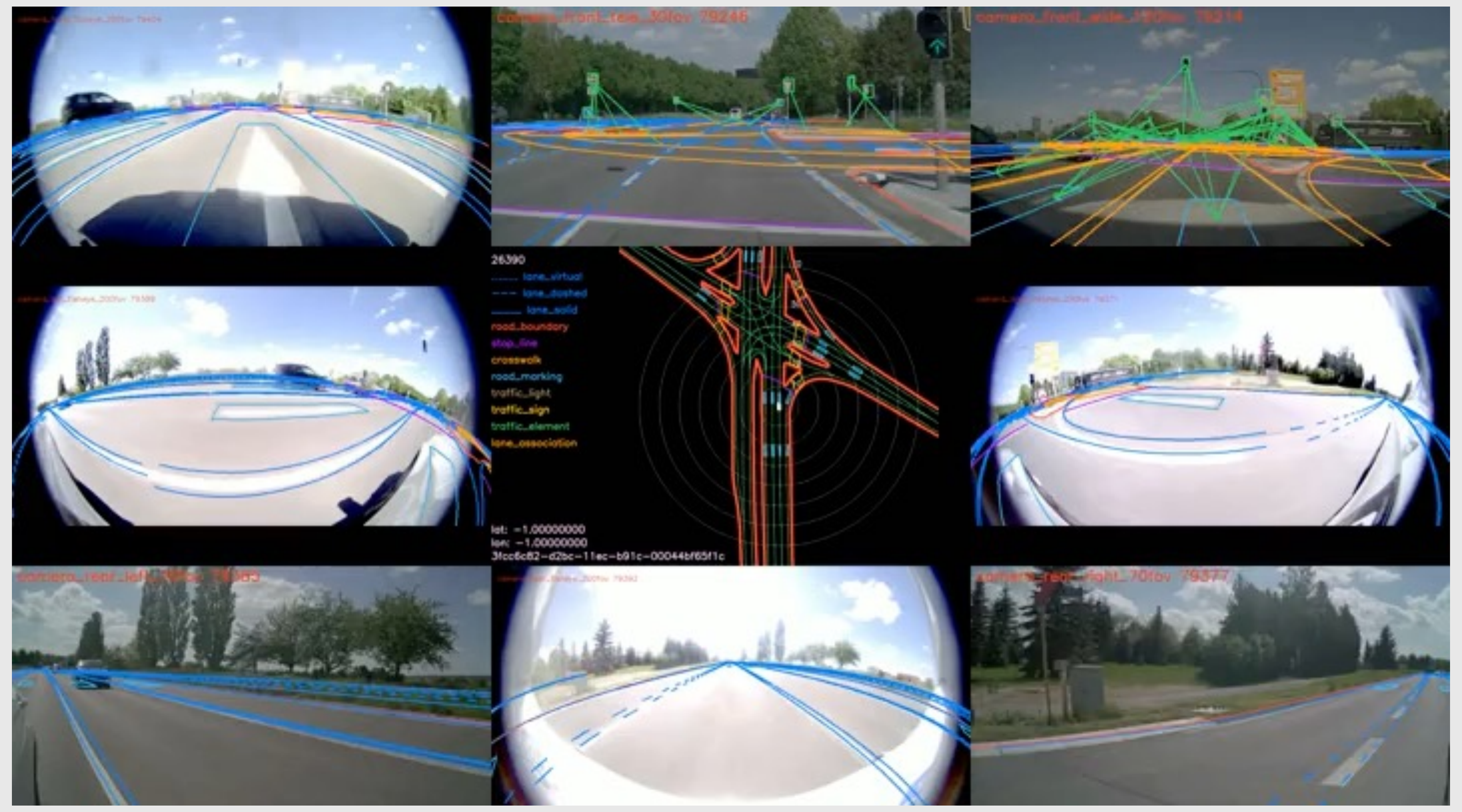
How Can We Use AV FMs?

Offline Processes

Autolabeling

Simulation

On-Vehicle AV Stack



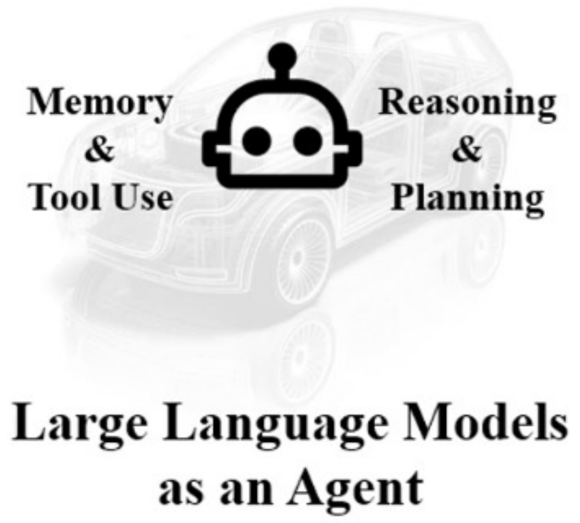
Novel End-to-End Architectures

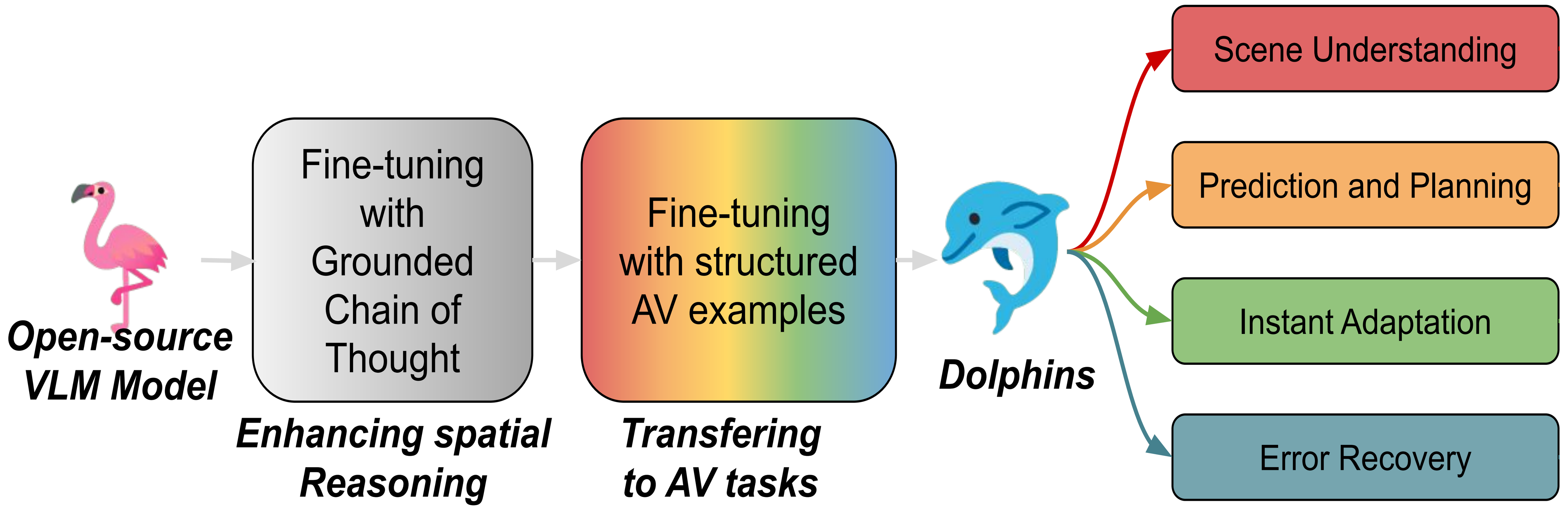
In-Cabin Assistance

Can FMs Drive?

AgentDriver and Dolphins as initial explorations

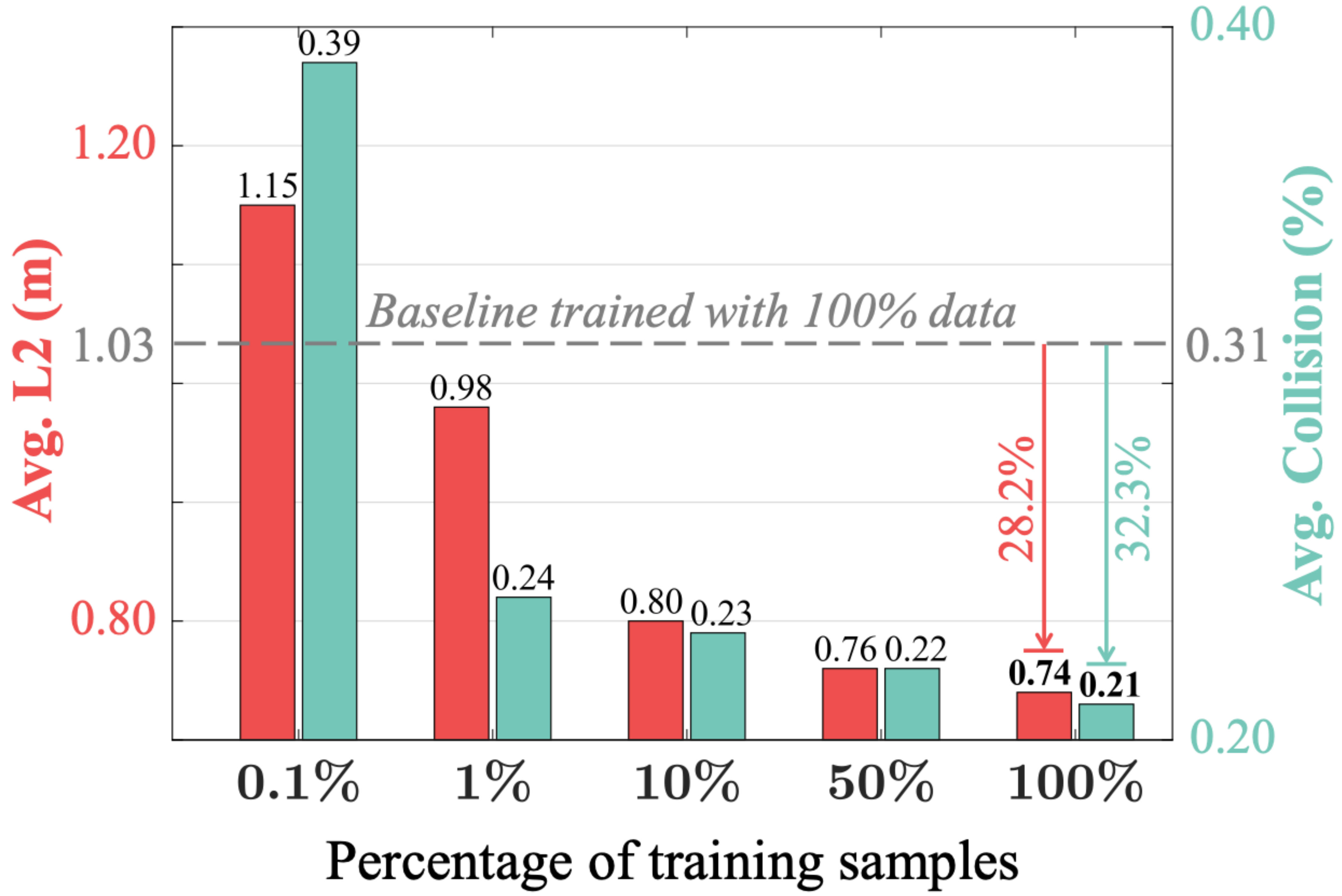
A Language Agent for Autonomous Driving

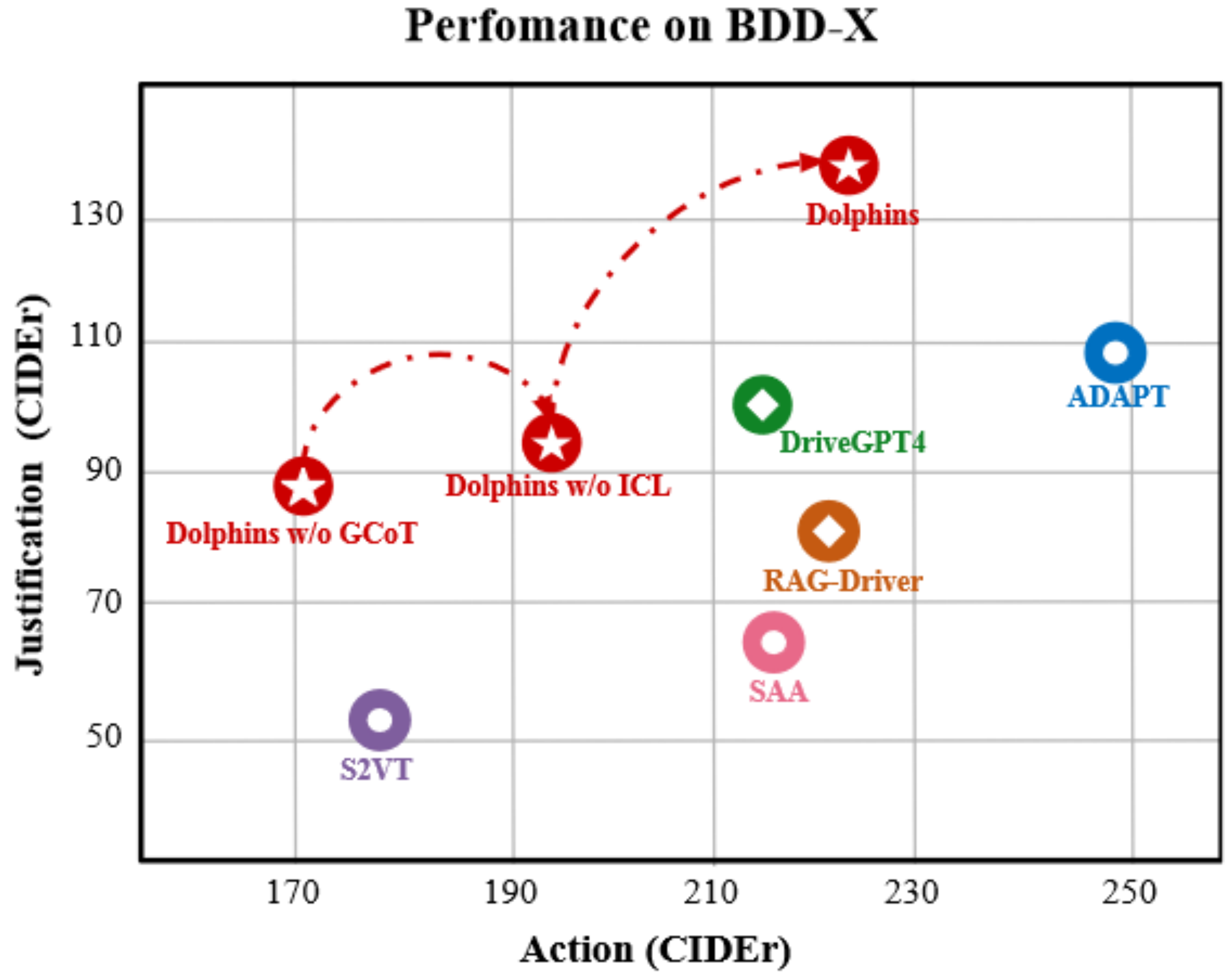




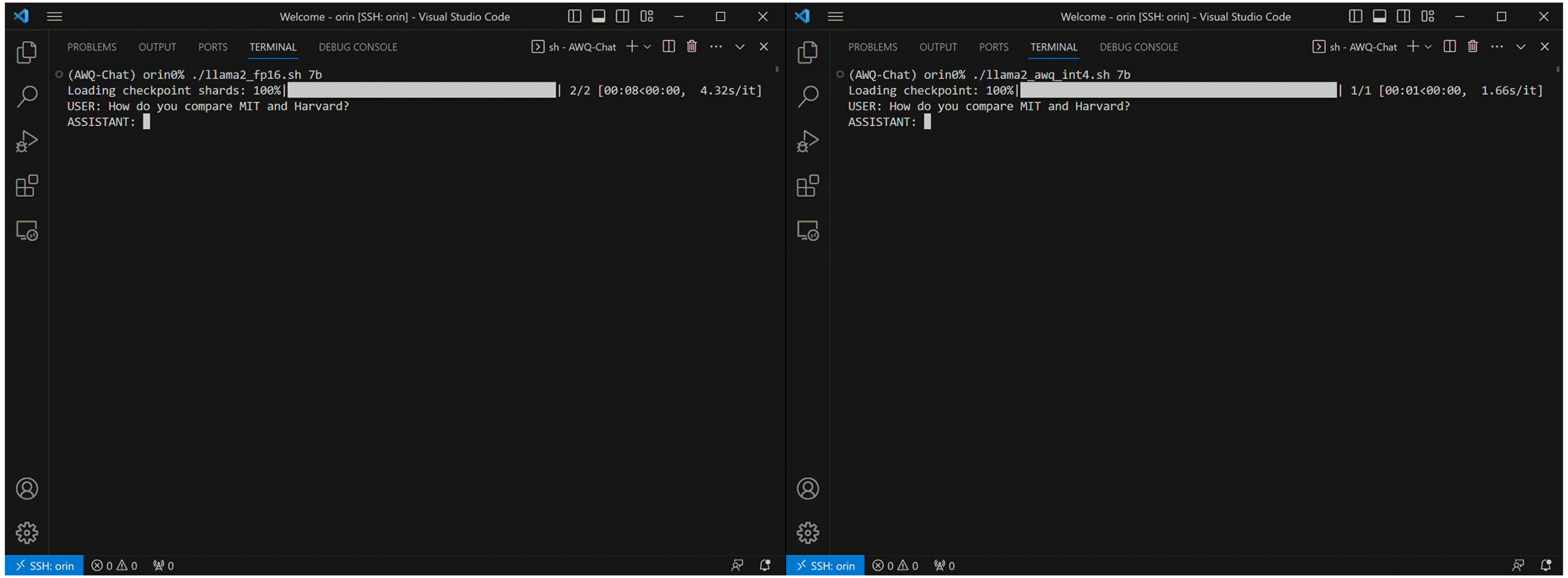
Mao*, Ye*, Qian, Pavone, Wang, *A Language Agent for Autonomous Driving.* Submitted.
https://usc-gvl.github.io/Agent-Driver/

Ma, Cao, Sun, Pavone, Xiao, *Dolphins: Multimodal Language Model for Driving.* Submitted.
https://vlm-driver.github.io/

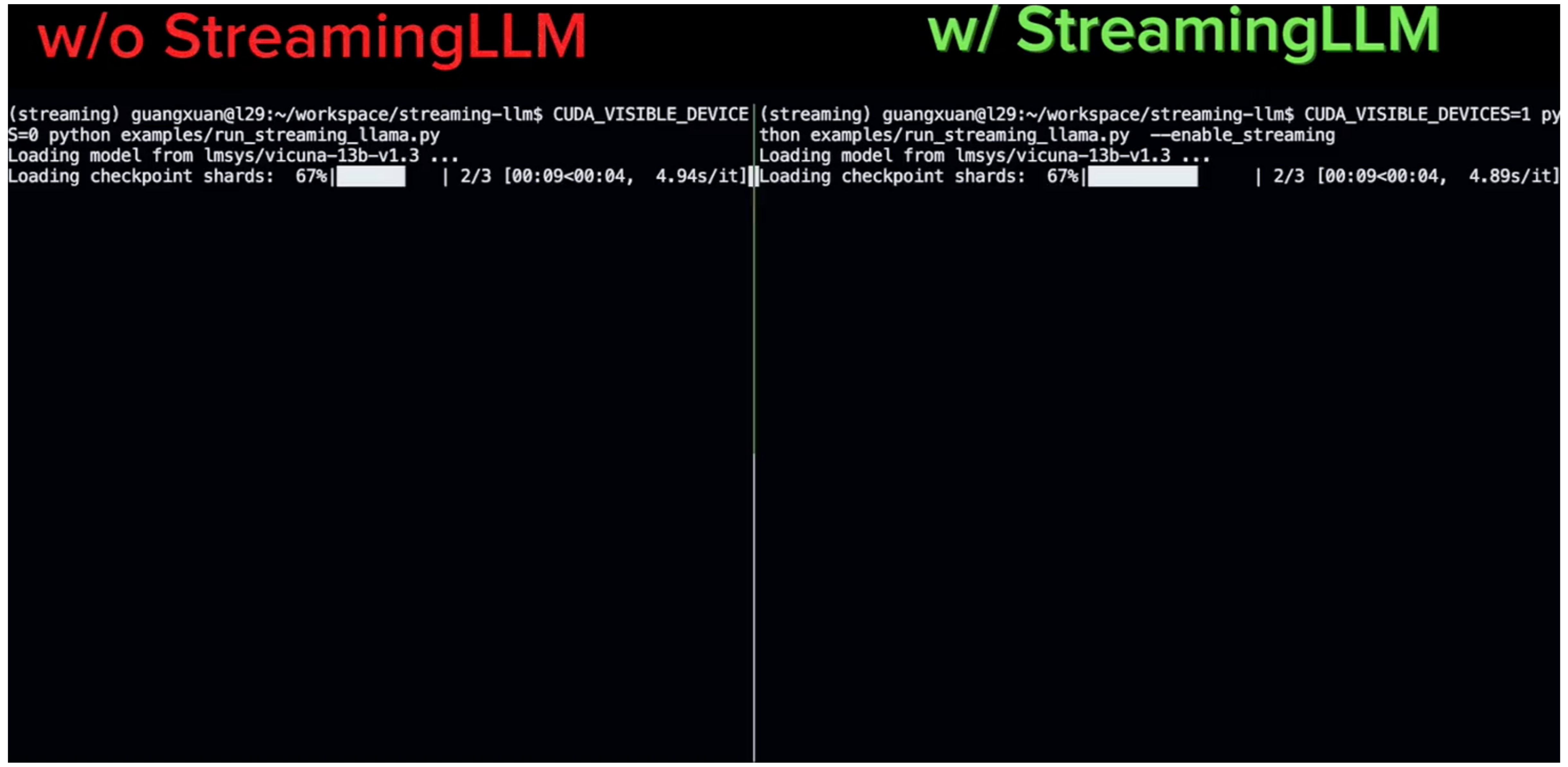




Can LLMs Drive *Practically*? Potentially!

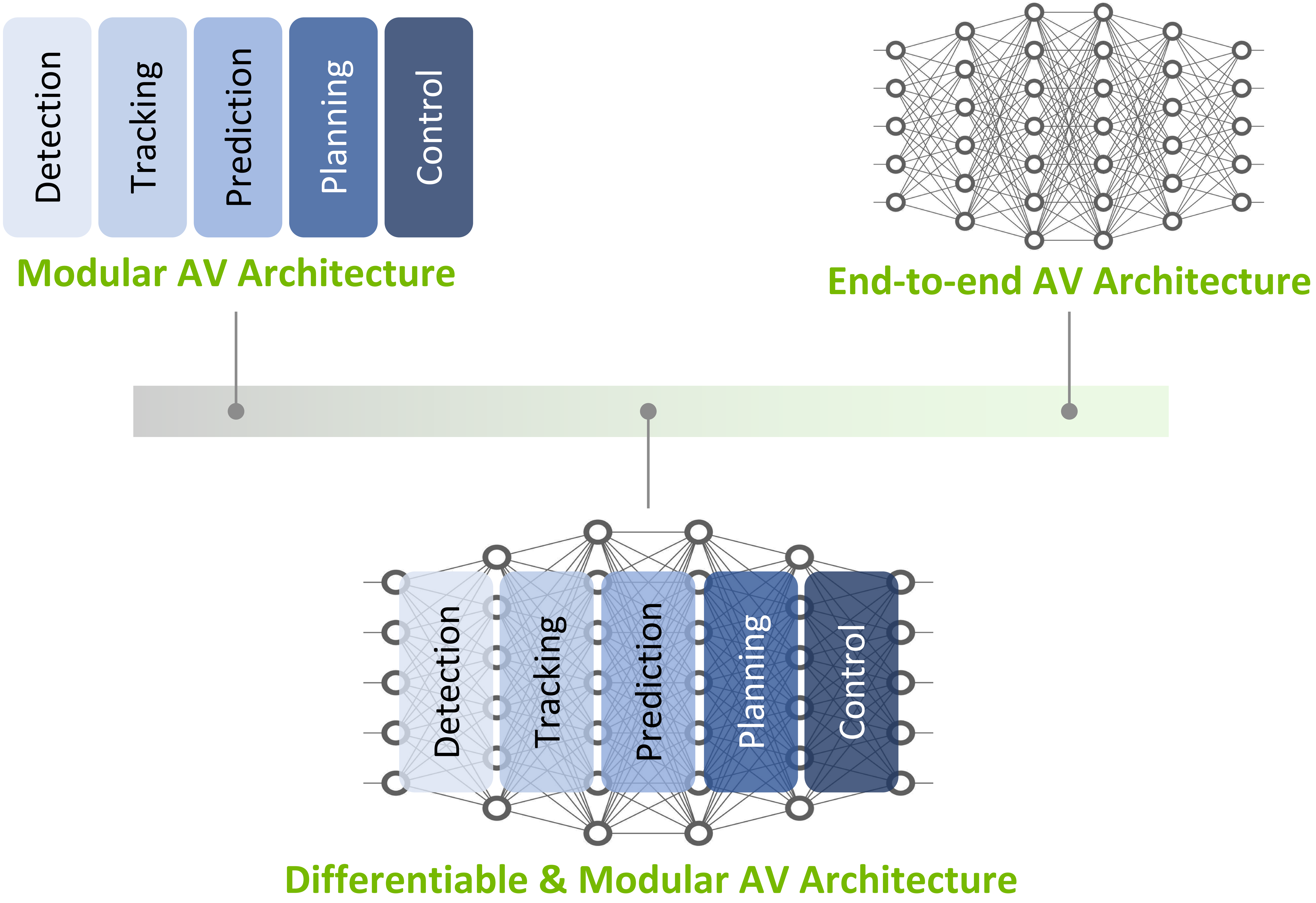


Lin, Tang, Tang, Yang, Dang, Han, *Activation-aware Weight Quantization for LLM Compression and Acceleration*, arXiv 2023

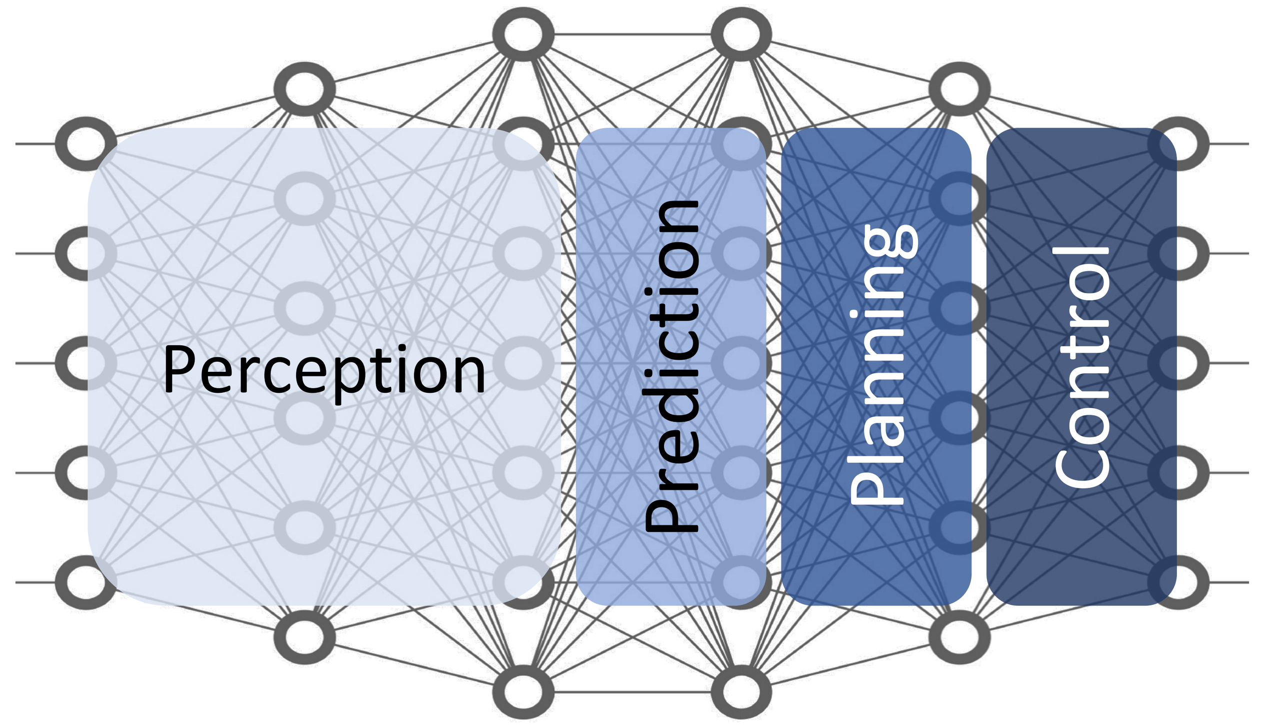


Xiao, Tian, Chen, Han, Lewis, *Efficient Streaming Language Models with Attention Sinks*, arXiv 2023

On System Architecting

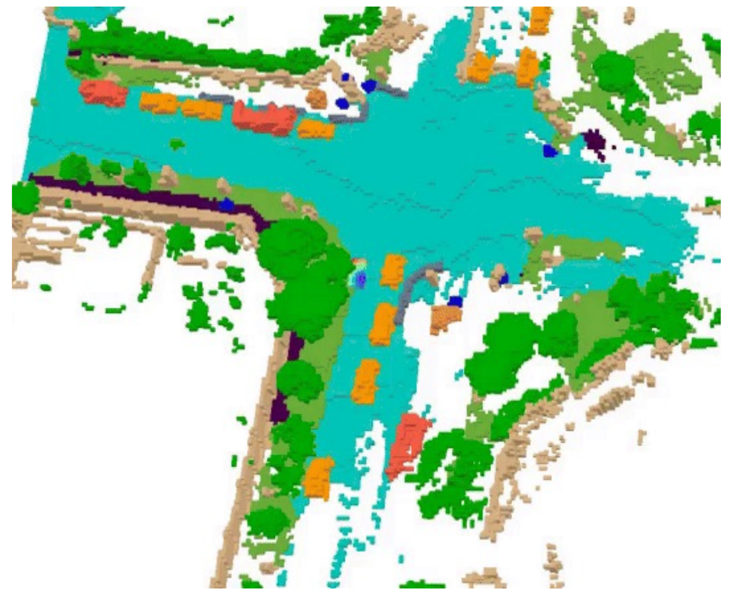


The Design Space is Extremely Large!



Differentiable & Modular AV Architecture

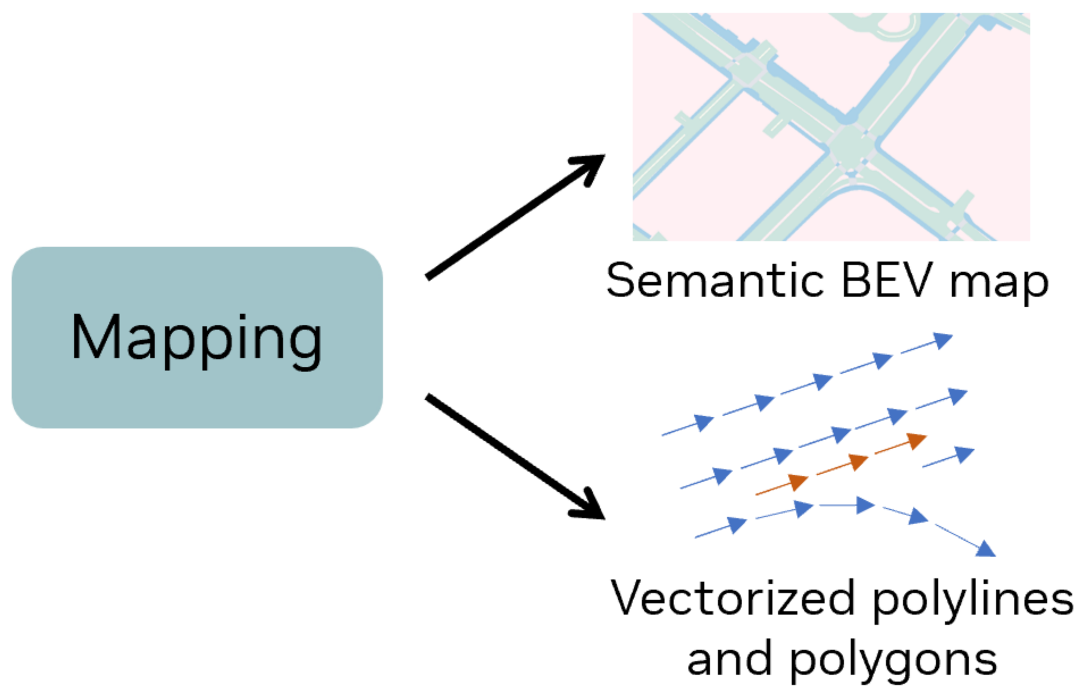
Choice of modules



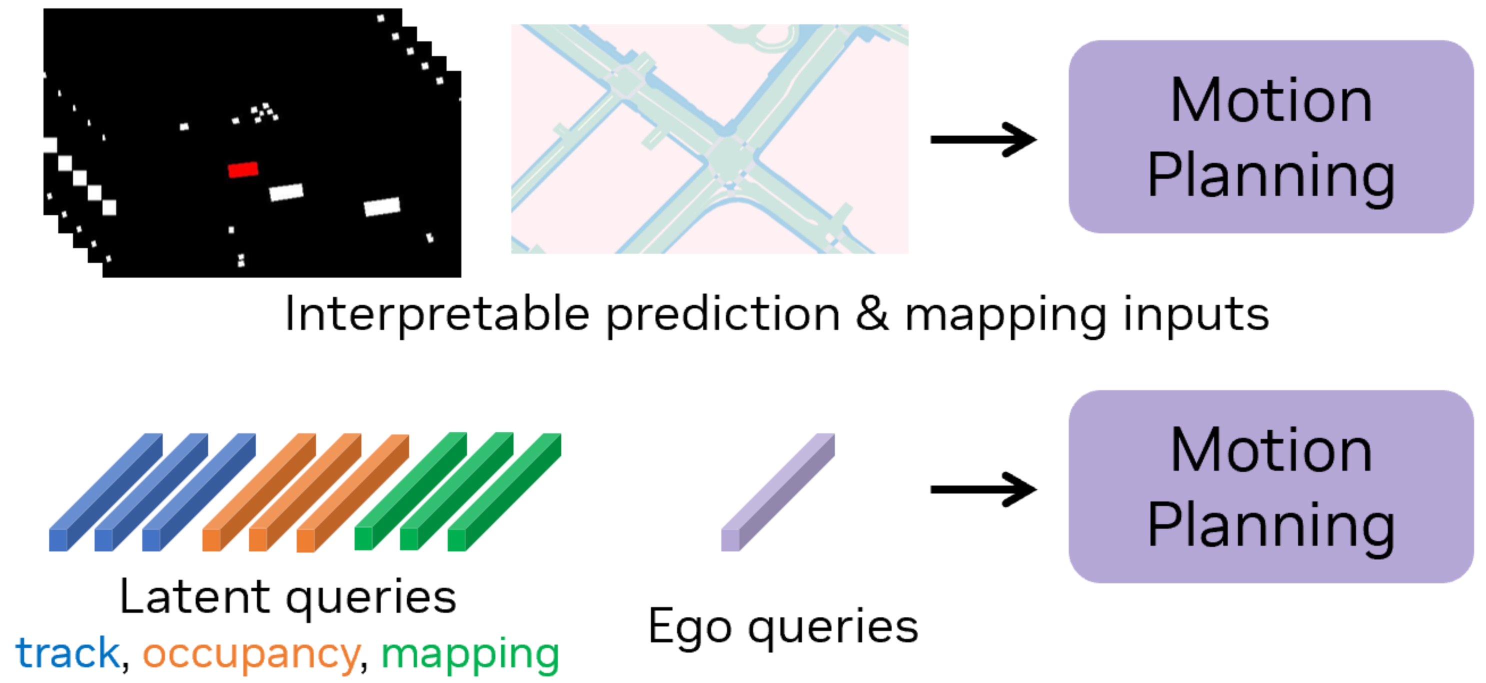
3D semantic occupancy network
OccNet (ICCV ‘23)

BEV occupancy flow & trajectory prediction
UniAD (CVPR ’23 best paper)

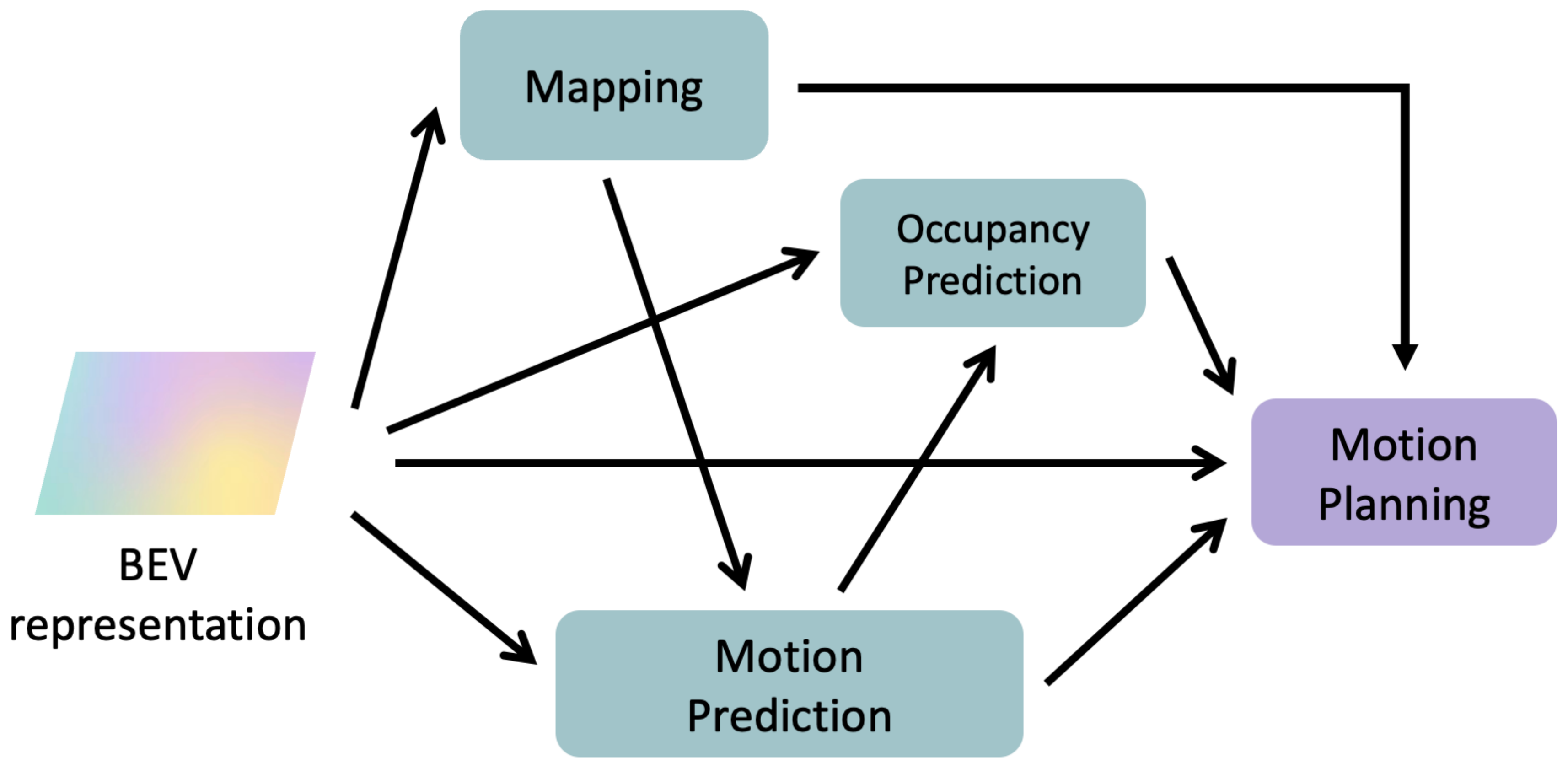
Choice of representations



Output representations



Input representations

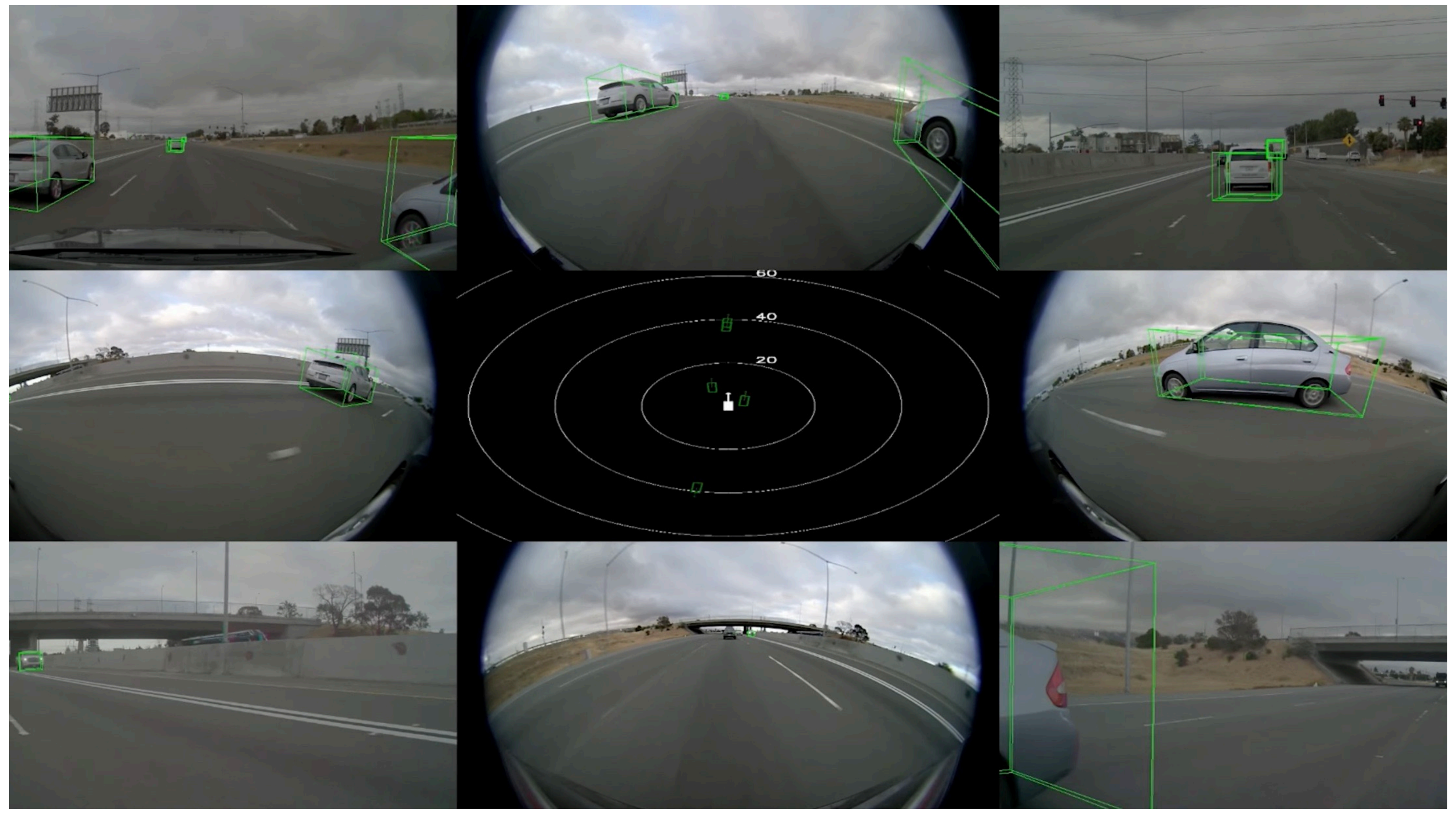


Coupled through module placement!
Compounded complexity

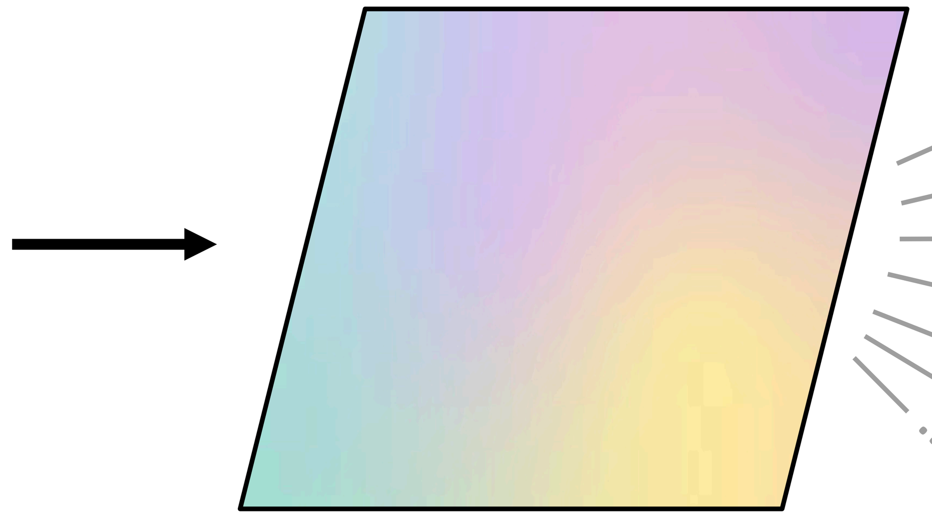
Weng, Ivanovic, Wang, Wang, Pavone, *PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving*, CVPR 2024

Building A Flexible Computational Driving Graph to Explore the Design Space

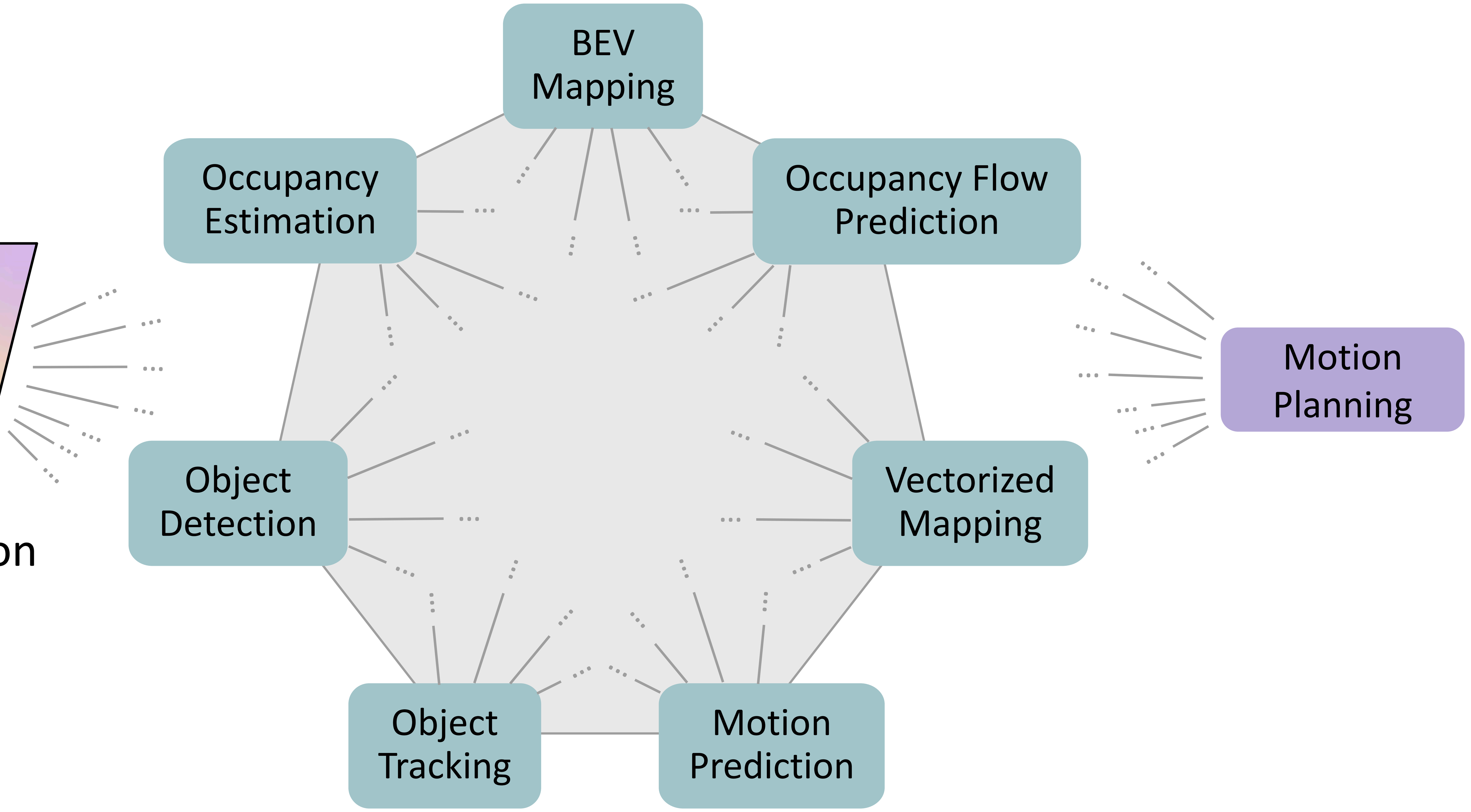
- **Necessity:** Which tasks/modules are essential for driving? Is there redundancy?
- **Placement:** How should modules be arranged? Sequentially? In Parallel? Hybrid?
- **Representation:** Should we use latent features (e.g., Transformer queries?), interpretable outputs (e.g., bounding boxes or BEV outputs), or a combination of both?



Multi-View Images



BEV Representation



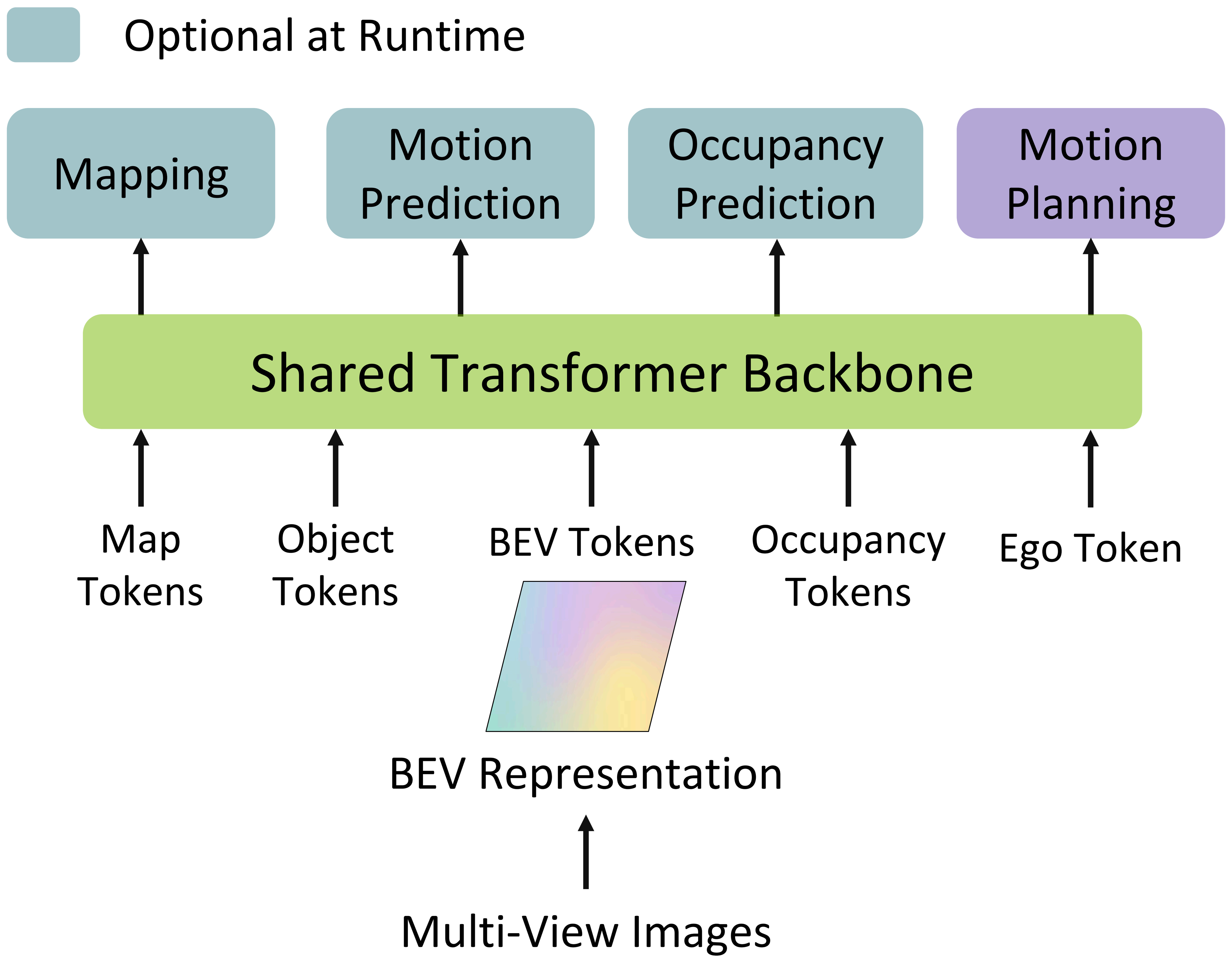
Fully-Connected Computational Driving Graph

Weng, Ivanovic, Wang, Wang, Pavone, *PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving*, CVPR 2024

PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving

Incorporating these insights yields a new state-of-the-art parallel architecture for end-to-end AV

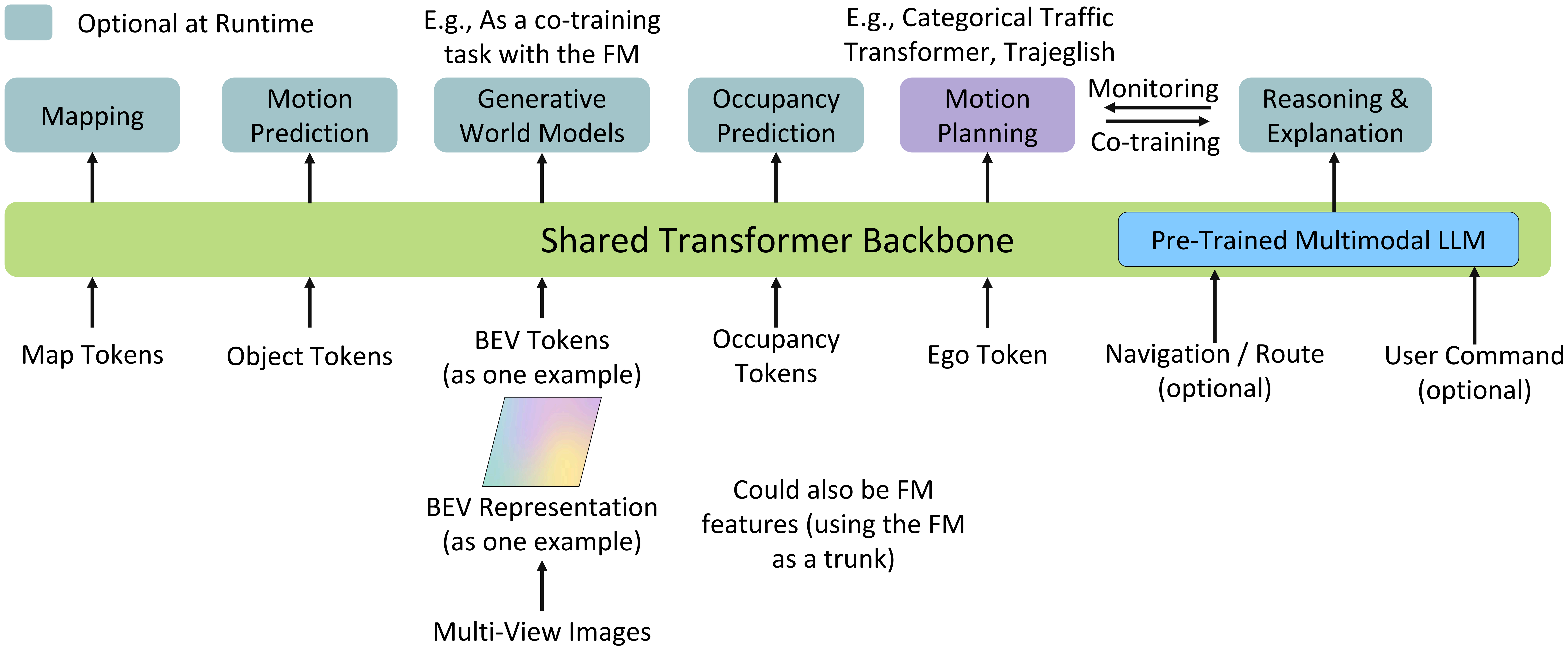
- PARA-Drive can run **4x** faster than state-of-the-art academic models (UniAD, CVPR 2023 Best Paper), while outperforming them in open-loop planning metrics and auxiliary tasks (e.g., mapping)



Weng, Ivanovic, Wang, Wang, Pavone, *PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving*, CVPR 2024

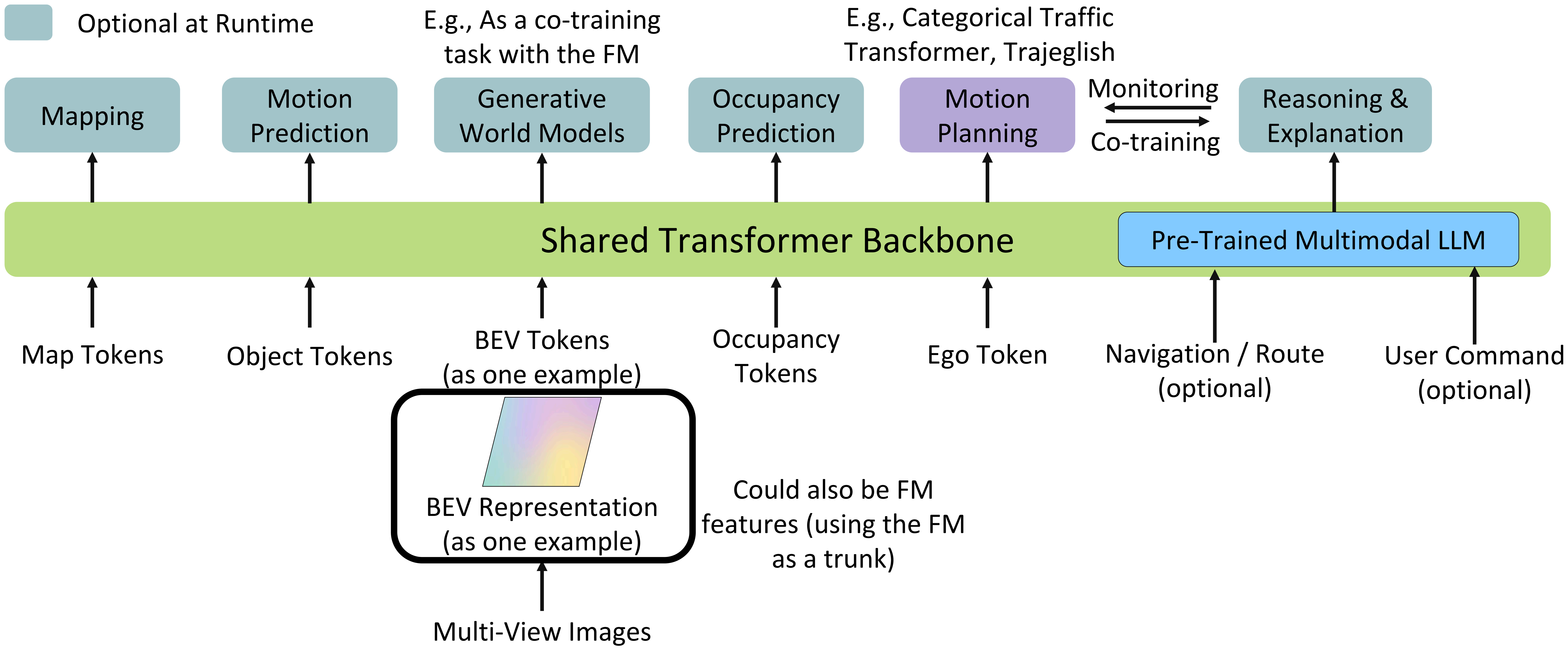
Supercharging End-to-End Driving with Foundation Model Features

- A parallel architecture with a shared backbone serves as a strong general archetype
  - Flexible with respect to input/output representations (which can differ at train and test time)
  - Supports training with multiple tasks (to shape internal features)
  - Can activate different decoder heads as desired, optionally enabling running FMs on-demand (e.g., in-cabin assistance) or at low frequencies (e.g., as a high-level planner or run-time monitor) online via distillation



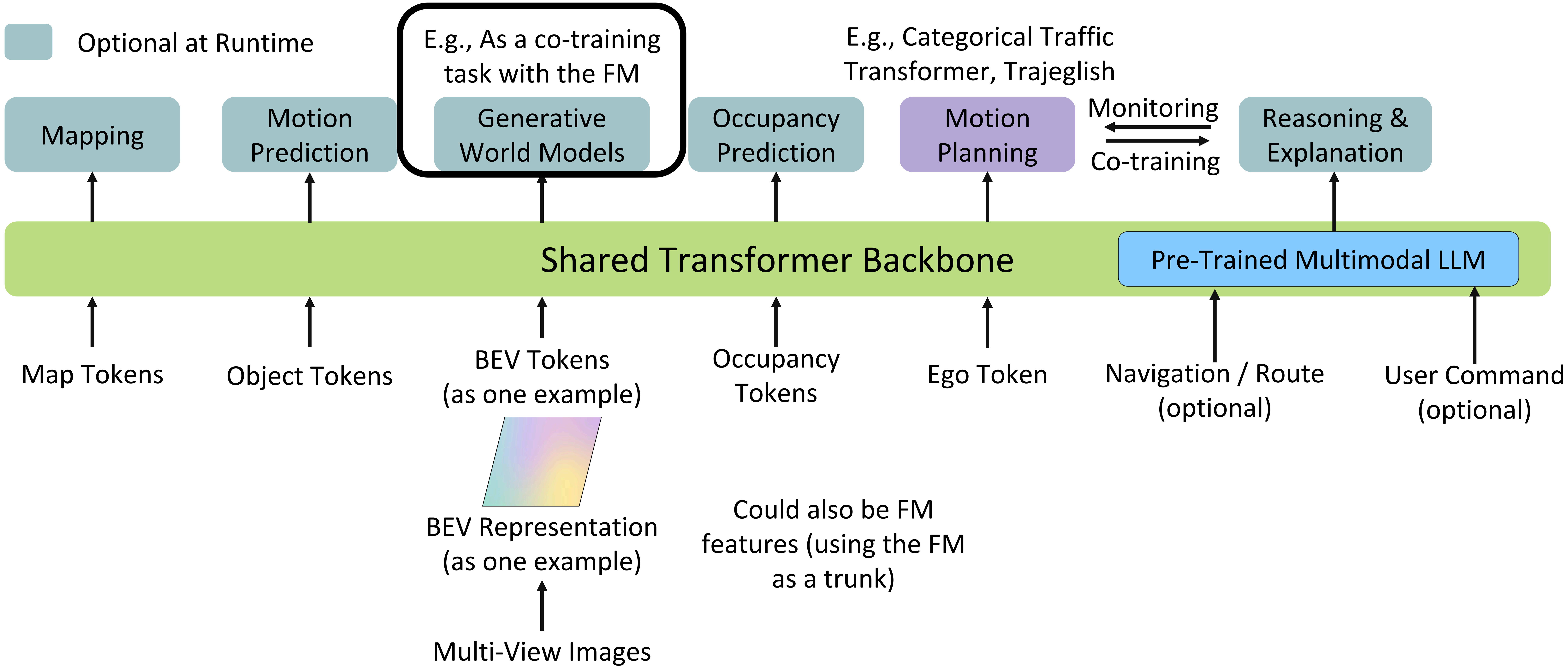
Supercharging End-to-End Driving with Foundation Model Features

- A parallel architecture with a shared backbone serves as a strong general archetype.
  - Flexible with respect to input/output representations (which can differ at train and test time)
  - Supports training with multiple tasks (to shape internal features)
  - Can activate different decoder heads as desired, optionally enabling running FMs on-demand (e.g., in-cabin assistance) or at low frequencies (e.g., as a high-level planner or run-time monitor) online via distillation



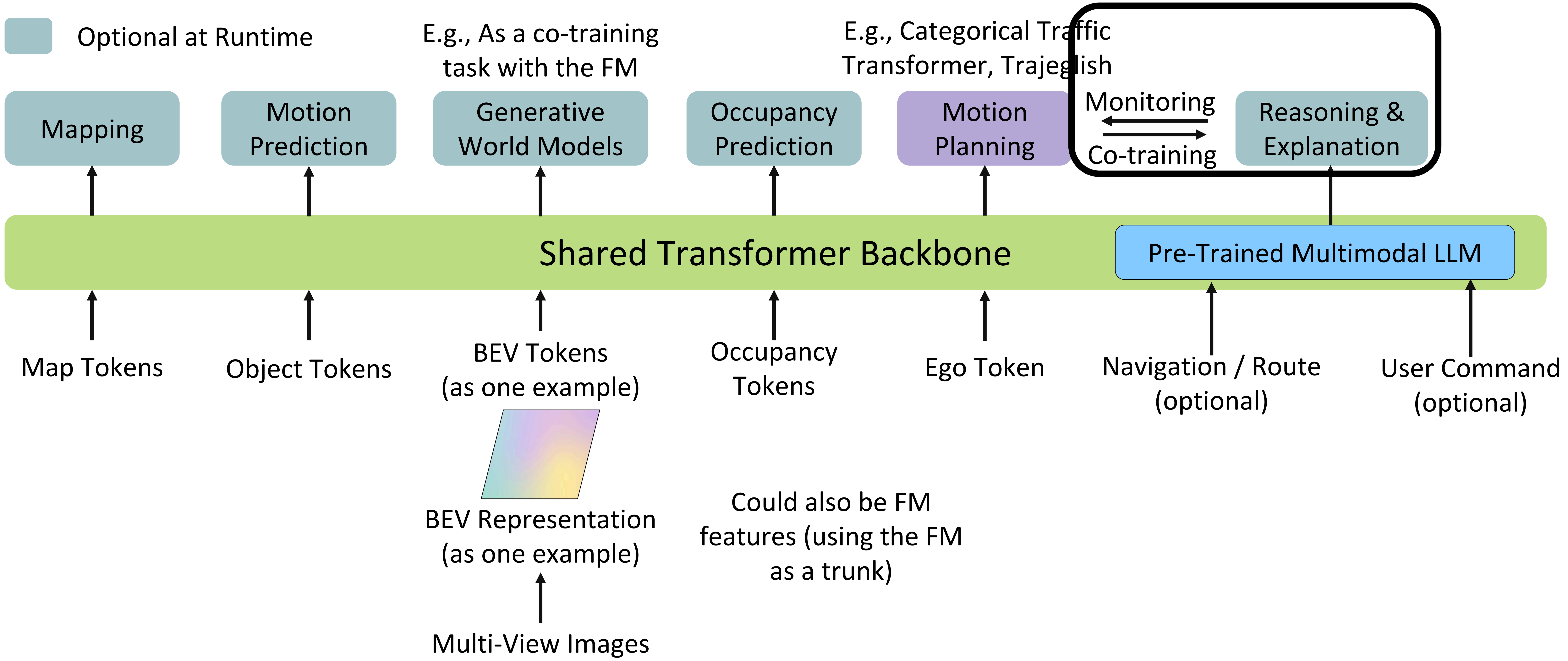
Supercharging End-to-End Driving with Foundation Model Features

- A parallel architecture with a shared backbone serves as a strong general archetype.
  - Flexible with respect to input/output representations (which can differ at train and test time)
  - Supports training with multiple tasks (to shape internal features)
  - Can activate different decoder heads as desired, optionally enabling running FMs on-demand (e.g., in-cabin assistance) or at low frequencies (e.g., as a high-level planner or run-time monitor) online via distillation



Supercharging End-to-End Driving with Foundation Model Features

- A parallel architecture with a shared backbone serves as a strong general archetype.
  - Flexible with respect to input/output representations (which can differ at train and test time)
  - Supports training with multiple tasks (to shape internal features)
  - Can activate different decoder heads as desired, optionally enabling running FMs on-demand (e.g., in-cabin assistance) or at low frequencies (e.g., as a high-level planner or run-time monitor) online via distillation



Tian, Li, Chen, Weng, Wang, Ivanovic, Pavone, *TOKEN: A Multi-modal Large Language Model with Tokenized Object-level Knowledge for Autonomous Driving* (in preparation)

Tian, Li, Chen, Weng, Wang, Ivanovic, Pavone, *TOKEN: A Multi-modal Large Language Model with Tokenized Object-level Knowledge for Autonomous Driving* (in preparation)

Conclusions

Key Takeaways

- FMs bring access to new data and capabilities that provide a quantum leap in long-tail generalization for AVs
- FMs have potential to empower the full AV program, from offline processes all the way to the online AV stack
- VFMs and MM-LLMs are emerging as two prominent FMs for AV - opportunities abound wrt specialization and unification
- FMs can be used to replace existing pipelines as well as to improve them
- FM now make closed-loop sim eval and training arguably within reach
- A parallel architecture provides key opportunities to embed FMs within a stack, while avoiding main drawbacks (e.g., enabling fast-slow reasoning pipelines)
- FMs are not black magic: strategies exist to confidently deploy them...
- ...at the same time, FMs provide key opportunities to *improve* the safety of AVs (e.g., via semantic run-time monitors)

More Information / Links to Papers

bivanovic@nvidia.com

borisivanovic.com

@iamborisi

NVIDIA

Autonomous Vehicle Research Group

+ VFM-AV Team

research.nvidia.com/labs/avg

github.com/NVlabs

NVIDIA