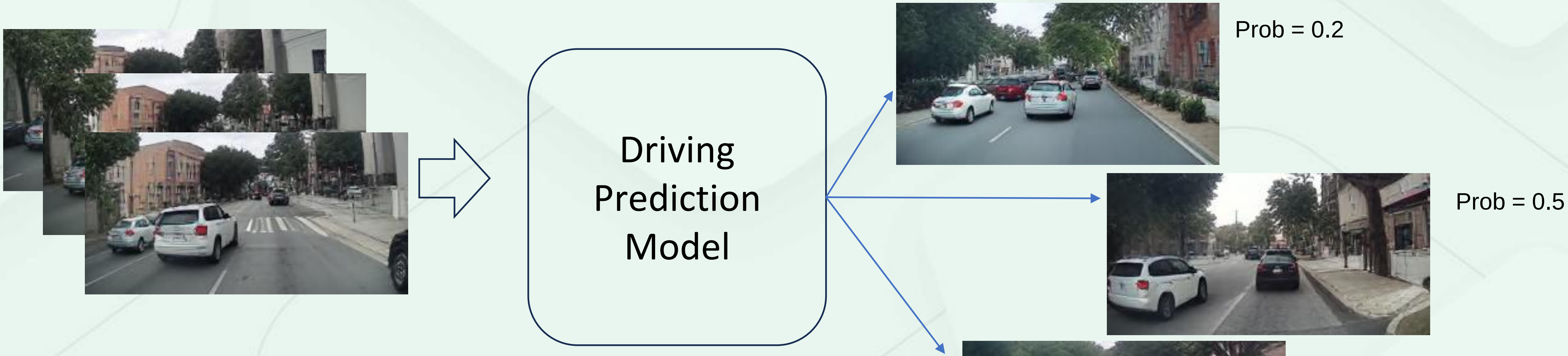
netradyne

A Foundational Driving Model Trained using Billions of Miles of Real-World Driving Data

David Julian
CTO, Netradyne Inc.

Generative Driving Model Potential



- For Drivers Today – Next Gen ADAS
  - Generalized Collison Warning: Pedestrian and driver intent
  - Generalized Unsafe Driving Warning: Running lights/stop signs/cutting off/…
- For AVs – Long Tail Data Challenge
  - Path planning incorporating generalized understanding

How can we create such a predictor?

- **Inspiration** – Foundational LLMs:
  - Trained on a large corpus of data using self-supervised training on next token prediction.
  - Exhibit emergent capabilities and generalization not explicitly trained for.

- **Our Approach** – Train foundational driving model on billions of miles of real-world driving data.
  - Using similar ideas
  - Multi-modal Data Types: Video, IMU, GPS, Vehicle Data, and AI Event Detections
  - Create ability to control the Ego vehicle.
  - Leveraging NVIDIA A100 GPUs using NCCL for distributed training.

How do we get billions of miles of real-world driving data?

Leverage Netradyne Driving Data

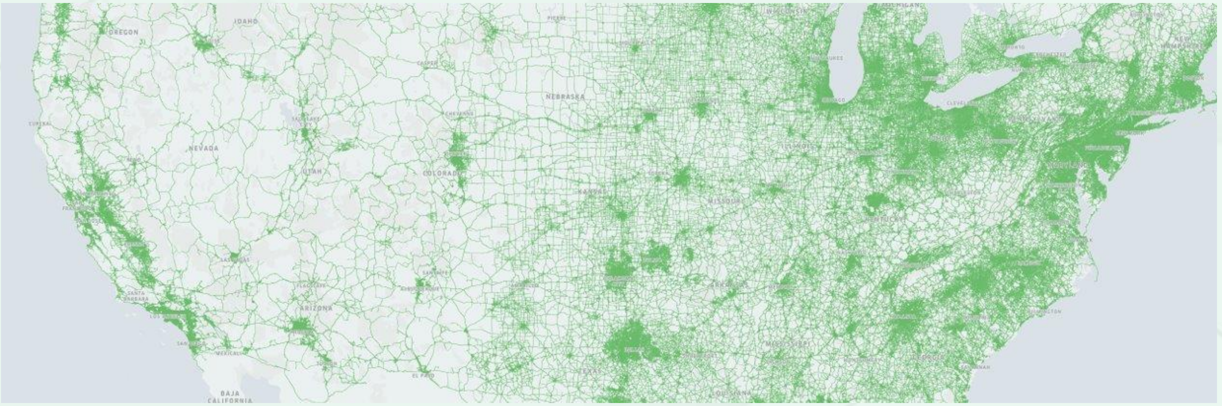
13B+
CUMULATIVE HIGH-RES MILES ANALYZED

500M+
Addl Miles/Month ANALYZED

2B+
CUMULATIVE EVENTS DETECTED

Rich high def, fully analyzed/categorized driving data

- Driving scenarios in all weather conditions, road types, localities, vehicle categories, …
- Accidents (tens of thousands), near-miss incidents, construction zones, pedestrians, bicyclists, traffic light, stop sign, lane changes, and more



- In comparison, AV industry has limited miles, <50 million across companies. [1]

[1] AVIA data shows 44 million+ autonomous miles driven and outstanding safety record, 2024

A Foundational Driving Model

Evidence of Emergent understanding

- **Green: Context frames**
- **Red: Foundational model output frames**

- Outputs indicating world model has emergent understanding of road environment, including cars as objects, lane change predicted showing understanding, driving rules, etc.

Model Generalizing to India

Evidence of Generalization and in-context learning

Original Video

Ground Truth Video Through tokenizer

Red: Foundational Model Output

Red: Foundational Model Output

- Foundational driving model world model trained on US data generalizing to India, even though not trained on India data
- Generated consistent with ground truth. Understands that auto-rickshaw is a vehicle object even though has never seen an auto-rickshaw in training data.
- Video tokenizer has room for improved fidelity

Architecture – Driving World Model

- Leveraging **Transformer** Based Architecture
  - Multi-modal
  - Modalities time-synced
  - Separator token between time frames

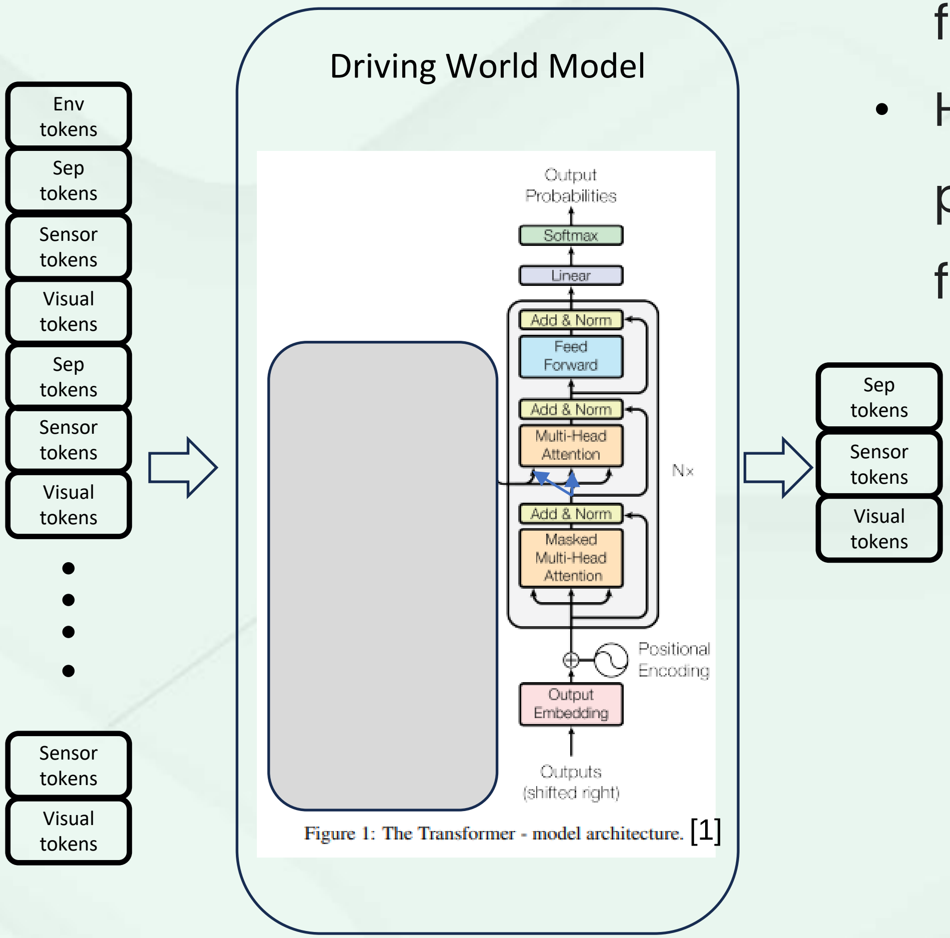


Figure 1: The Transformer - model architecture. [1]

- Output provides next frame prediction.
- Hidden states encapsulate probabilities over the future

[1] A. Viswani, et.al. “Attention is All You Need,” arXiv:1706.03762, 2017

Architecture – Tokenization Vocabulary

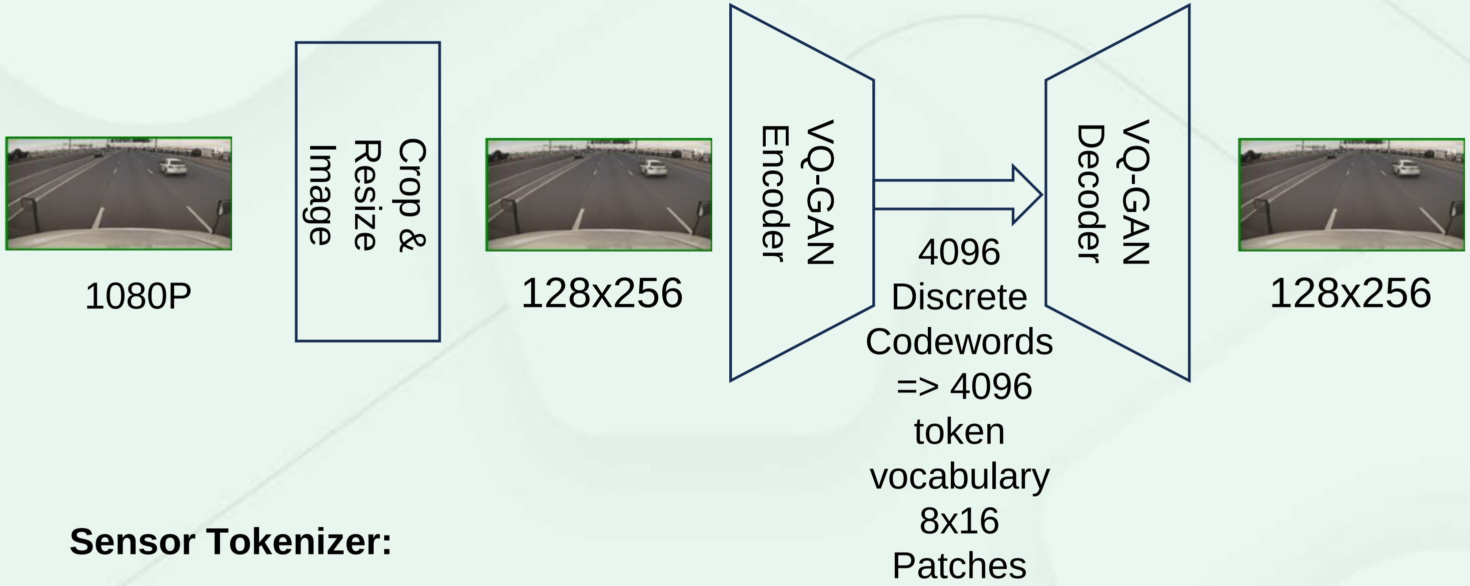
Environment Tokens:

- Ego Vehicle Class (class 1 to 8)
- Micro Weather (Clear, Rain, Fog, Snow)
- Time of day (Day, Dawn, Dusk, Night)

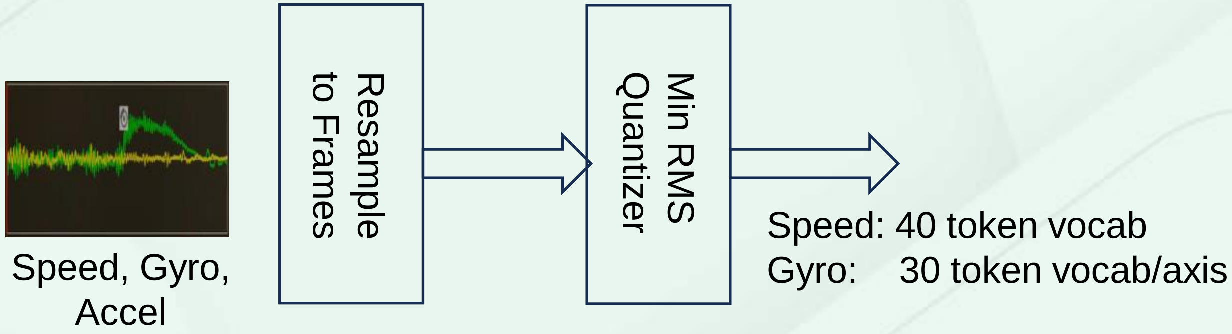
16 token vocab

Separator Tokens:
1 token vocab

Vision Tokenizer:

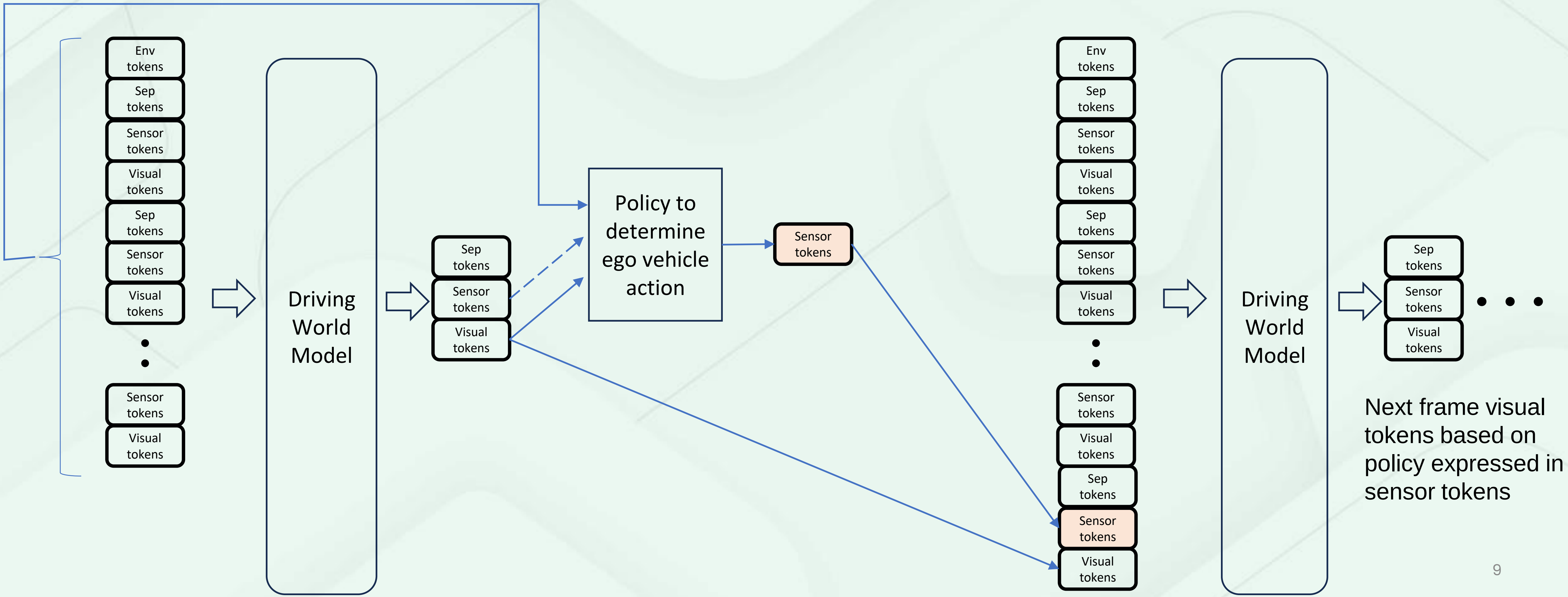


Sensor Tokenizer:



Controlling the Ego Vehicle – External Policy

- Use sensor tokens to control the Ego vehicle motion



netradyne

Example controlling ego vehicle

Forcing action at intersection

Ground Truth turns right

Control forces left turn

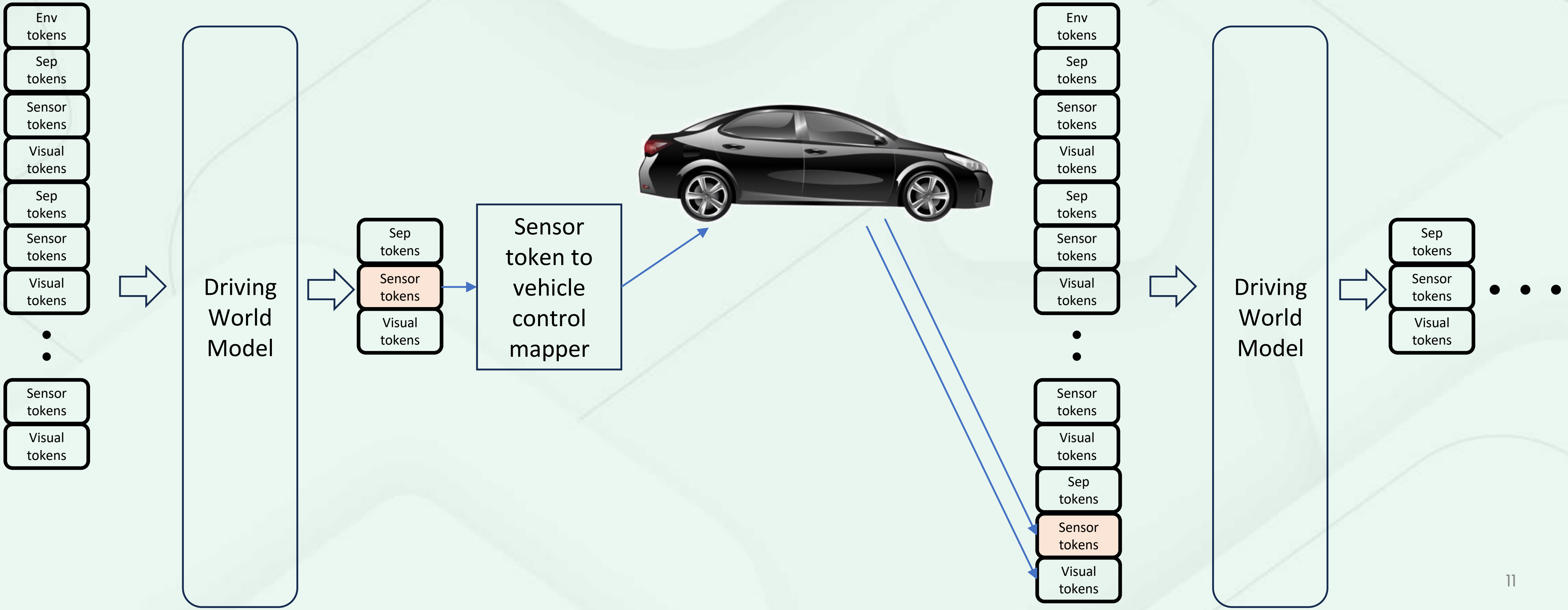
Control forces go straight

Control forces right turn after stop

- Foundational Driving Model enables ability for external policy model.
- Scenario modeling, beam searching, and path prediction evalution

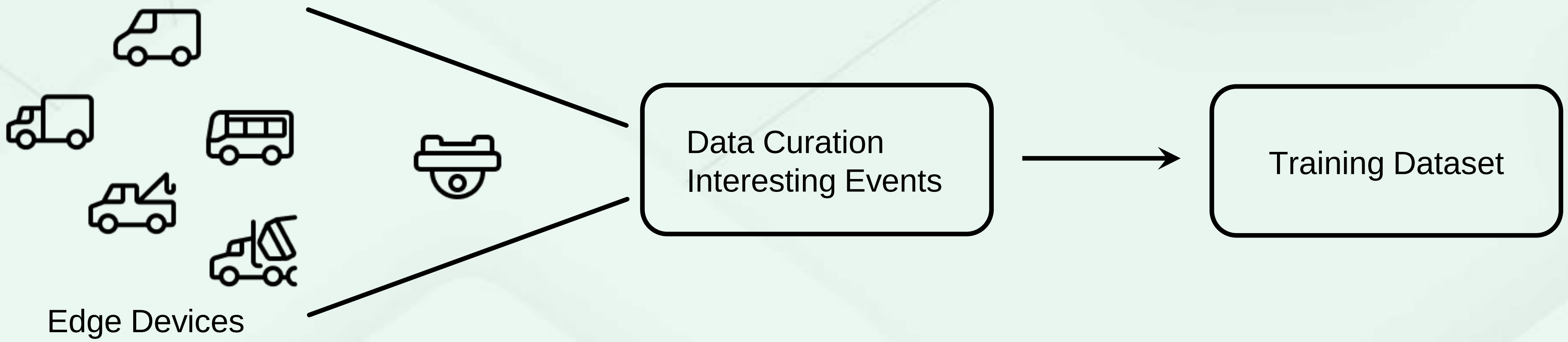
Concept: Driving World Model as a Policy

- Use sensor tokens to control the Ego vehicle motion on-line
- Align model to safe driving



Training Dataset Curation

- Targeting a smaller foundational model to run in real-time on-edge use cases
- Most driving is uninteresting/**mundane**. Interesting events are **sparse**.
  - For given model size risk using too much model capacity for fidelity of frequent mundane driving
  - Allocating insufficient capacity for interesting/important events.
- Motivation: Curated and cleaned data facilitate enhanced performance for LLMs [1].



[1] M. Javaheripi and S. Bubeck, “Phi-2: The surprising power of small language models”, Dec 12, 2023

Lane Change Hallucination

90M parameter model trained on curated data outperforms on rare events

- Left case, training included a lot of mundane driving data, starts hallucinating when forced to make a lane change,
    - Expect better pixel fidelity in common mundane driving
- Right case, trained on more curated data, completes the lane change
    - Use AI triggers to sub-select/curate and focus on interesting segments out of ten billion miles

netradyne

There are many rare events -> Billions of Miles









Rare Events - Accidents

Only the model trained on accidents could conceive of accidents

- Trained models without accidents data and with accident data
- **No Accident Model:**
  - Drove right through vehicle and kept driving, or
  - Hallucinated creation of a round-about and accident vehicle just continued around round about
- **Accident model:** Accurately predicted collisions and follow-on effects
- Example of importance of covering long tail events in training data with many examples per event

Summary and Next Steps

- Believe foundational driving model trained on billions of driving miles will be critical for AV2.0

- Built foundational driving model on data curated from billions of real-world miles
  - Showed architecture to create and control driving model
  - Showed examples of emergent abilities and generalization

- Showed real world data is important for long-tail of corner cases

Next Step:

- Align driving model policy to safe driving

- Add additional sensors, such as RADAR, multiple cameras, …

netradyne

Questions?