Valeo

SMART TECHNOLOGY
FOR SMARTER MOBILITY

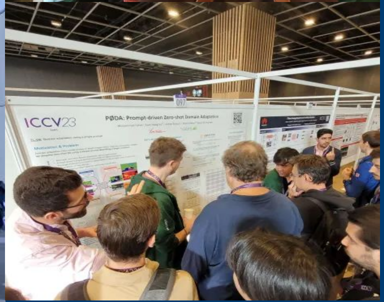
valeo.ai

better, clearer & safer
automotive AI

Foundation Models on Wheels

Large Scale Self-Supervised Learning for Autonomous Driving

Florent Bartoccioni



valeo.ai

better, clearer & safer automotive AI

https://valeoai.github.io/blog/

- ~30 people (researchers, PhDs)
- Dedicated to open research
- 13000+ citations
- 58 open-sourced codebases
- 4000+ stars on github
- 10's of academic partnerships across France and Europe

CVPR JUNE 17-21, 2024 SEATTLE, WA

NEURAL INFORMATION PROCESSING SYSTEMS

ICCV23 PARIS

ICRA2024 May 13-18, Yokohama, Japan

3DV 2024

Springer INTERNATINAL JOURNAL OF COMPUTER VISION

EPFL

Inria

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

École des Ponts ParisTech

SORBONNE UNIVERSITÉ

ENS DE LYON

Valeo

UGA Université Grenoble Alpes

TELECOM ParisTech

Université Bretagne Sud ubs

From ADAS* to AD**

Spectrum of Vehicle Automatization

Driving Assistance

- Blind spot detection
- Cruise control

Forward collision warning
+
autobrake

↓ **56%** Front-to-rear crashes with injuries

Lane departure warning

↓ **21%** Injury crashes

*ADAS = Advanced Driving Assitance Systems **AD = Autonomous driving

Insurance Institute for Highway Safety
Images from ISO.org

From ADAS* to AD**

Spectrum of Vehicle Automatization

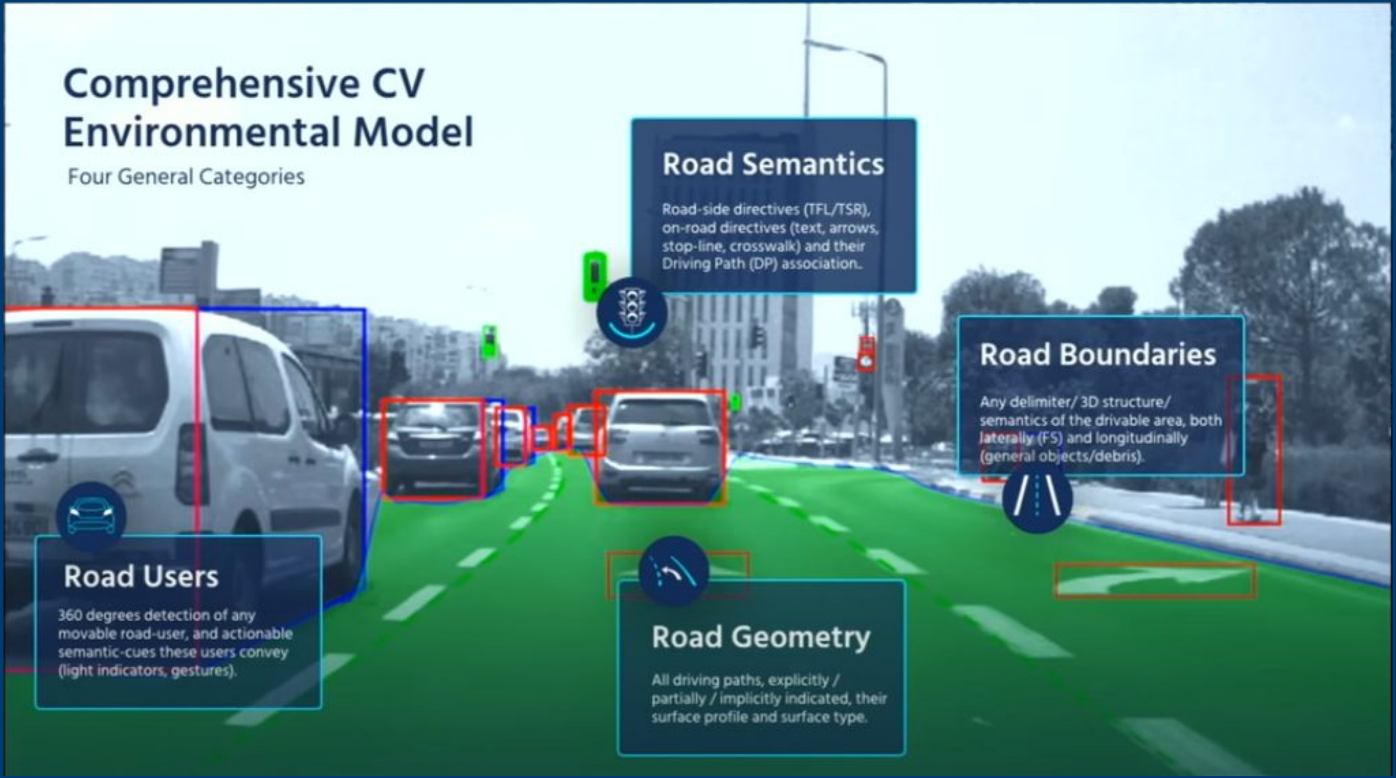
Driving Assistance	Limited Self-Driving	Full Self-Driving
▪ Blind spot detection ▪ Cruise control	▪ Parking valet ▪ Highway pilot	▪ Robot taxis ▪ Delivery vehicle

Towards safer, more efficient and more available mobility

*ADAS = Advanced Driving Assitance Systems   **AD = Autonomous driving

Core need of driving: representing the environment

Scene geometry, dynamic, semantic...



Adapted from https://youtu.be/HPWGFzqd7pI?t=1046

Core need of driving: representing the environment

How ?

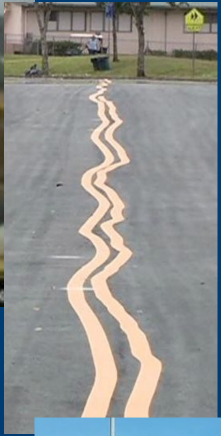
Ontology

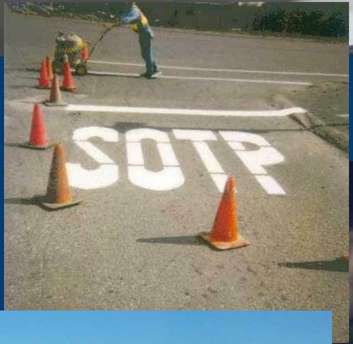
- Explicitly represent everything
- Detection, Segmentation
- Powered by human annotation
- What is not defined does not exist

Adapted from https://youtu.be/fKXztwtXaGo

Importance of learning at scale

The world is full of edge-cases

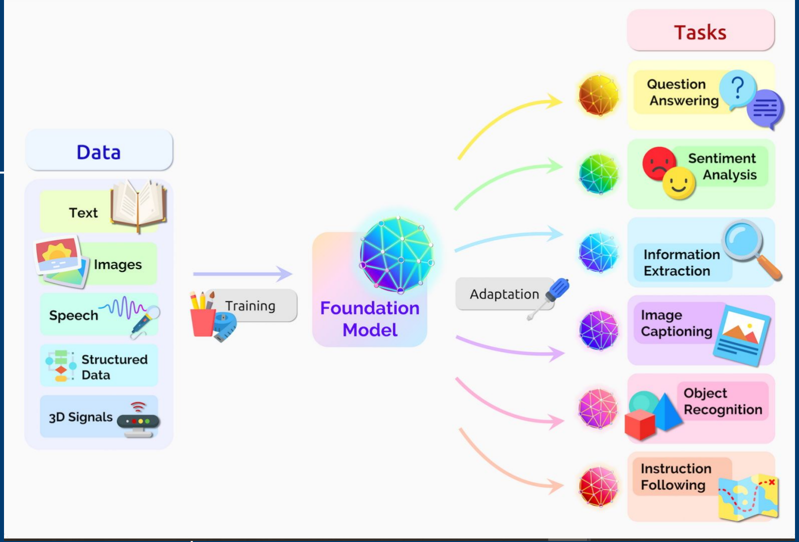




Learning at scale -> Foundation model

A little bit of vocabulary definition

Bommasani et al., On the Opportunities and Risks of Foundation Models, arxiv 2021



What do we call Foundation model?

Training

- **Expensive** to train → **one-shot**
- Trained on unlabelled (or weakly-labelled) data
  - Huge/Large scale
- Big model (> 1B ?)

Usage

- General purpose AI → can be applied to a **wide range of use cases**
- Easy to derive another model more specialized
- Training-free / zero-shot / cheap **specialization**

Foundational Models Strategies

Distillation or Self-supervision (non–exclusive)

Distillation from third-party models (DINO, CLIP, LLaVa etc.)

+ Less costly
+ Piggy back on GAFAM's monstrous budgets

– Limited training on driving data ?
– Ownership ? IP ?
– Bias control ?

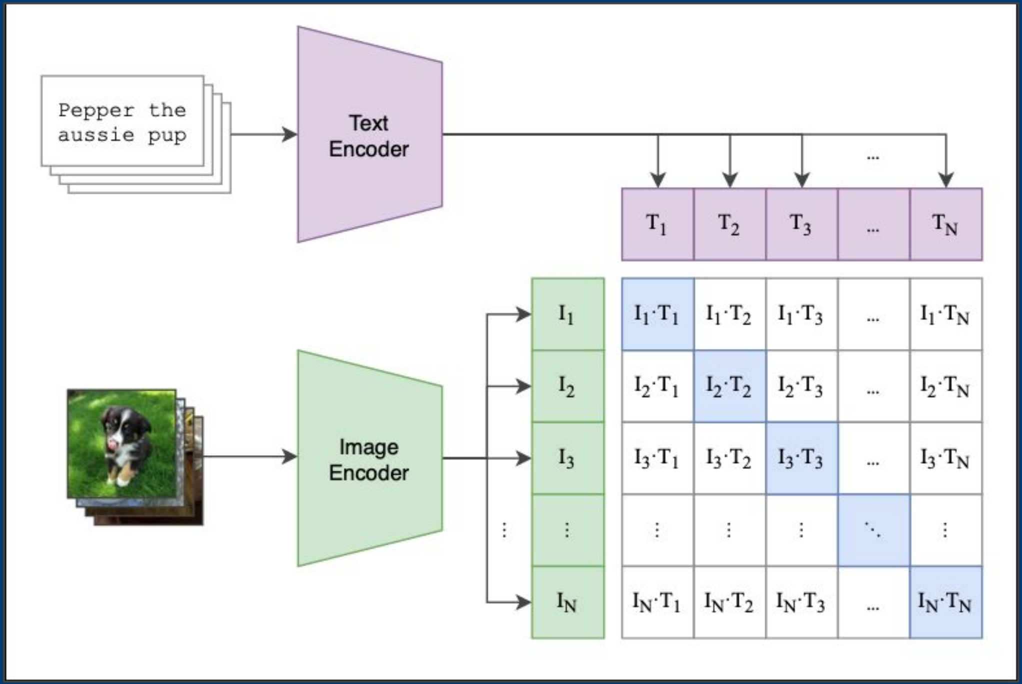
Learning from scratch with self-supervision

+ Entirely learned for the automotive domain
+ Adaptable to new domains (record+train)
+ Total control and ownership

– More costly
– Require robust data and infrastructure strategies

Distillation - Open-vocabulary, aligning text and images

CLIP : Learning Transferable Visual Models From Natural Language Supervision

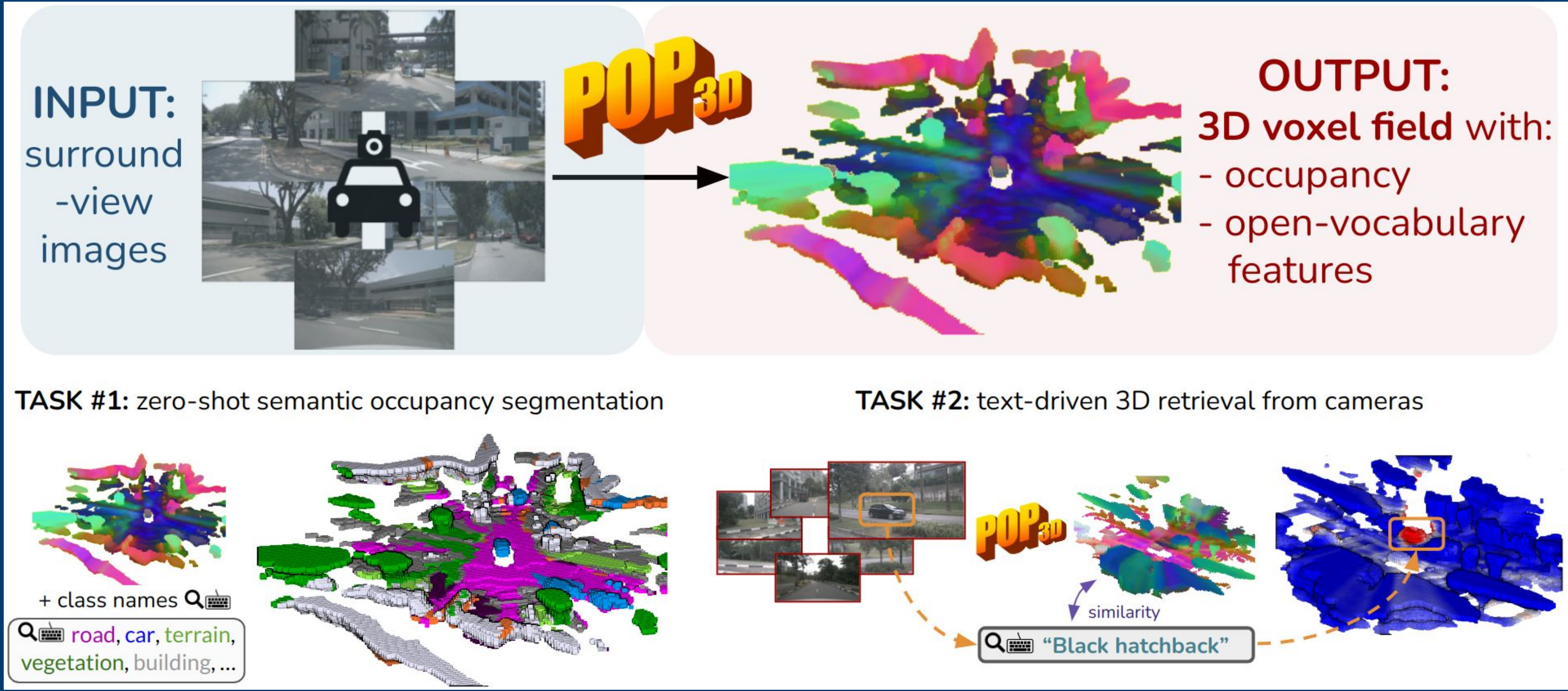


- Contrastive learning
  - Contrast positive/negative pairs
- Trained using 400 millions (image / text) pairs **extracted from internet**
  - Meta-data
  - Legends
- Align **perception <-> language**

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

Distillation - Geometry + CLIP

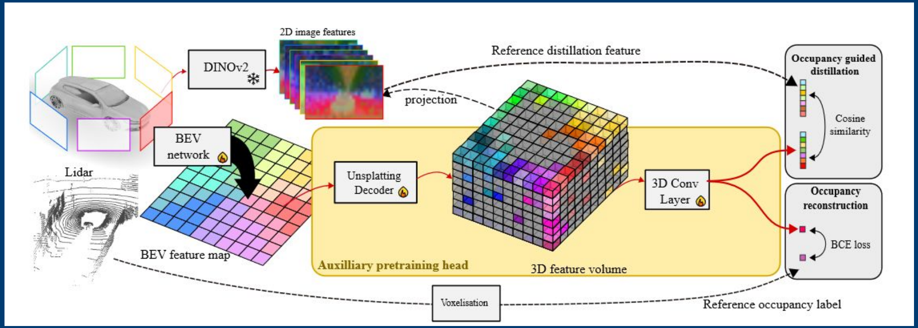
POP-3D: Open-Vocabulary 3D Occupancy Prediction from Images (NeurIPS 2023) [v.ai]



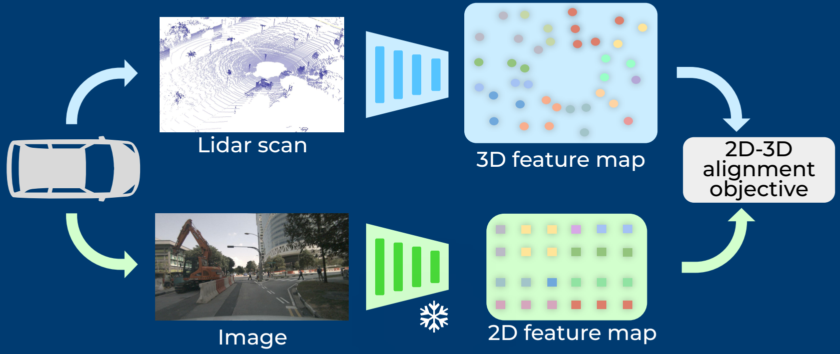
Distillation - Other works from the team

LiDAR, Camera, Camera+LiDAR

OccFeat
Self-supervised Occupancy Feature Prediction for Pretraining BEV Seg. Nets
(WAD CVPR 2024) [v.ai]



ScaLR
Three Pillars improving Vision Foundation Model Distillation for Lidar
(CVPR 2024) [v.ai]



- Explicit geometry as base
- 3D or BEV occupancy
- Features from foundation model
- No human annotation

How to learn foundational models for AD ?

The challenges

1. What data ?
2. What network architecture ?
3. What supervision ?
4. How to scale ?

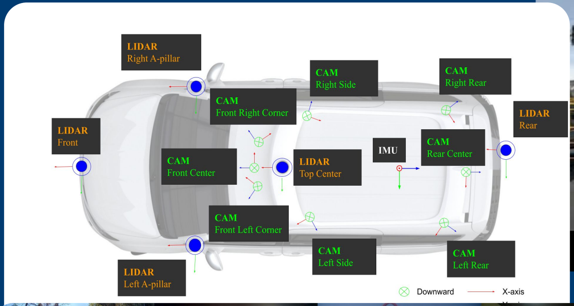
1/ Source of publicly available data

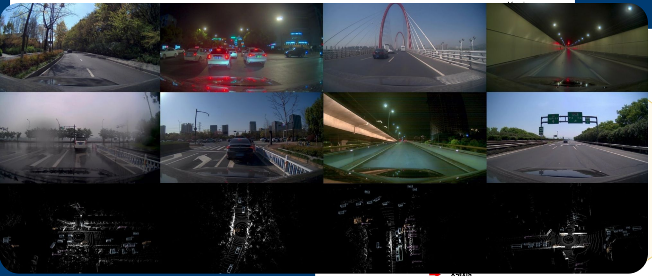
Heterogeneous situation

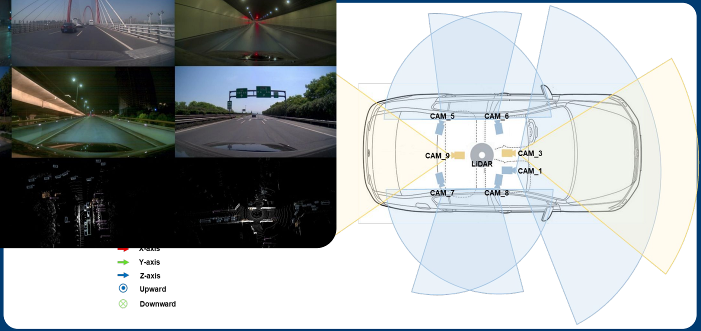
ONCE + Nuplan
(calibrated multi-cam + LiDAR)
~ 120h of driving data @10Hz

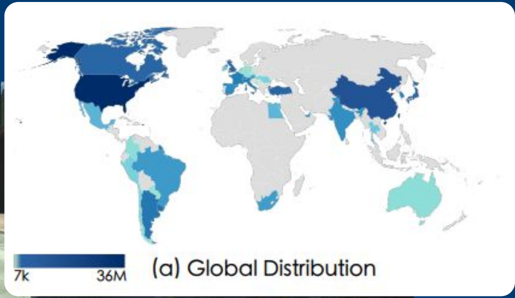
Includes rare events

OpenDV-Youtube
(only non-calib. front-cam)
~ 1700h of driving data @10Hz



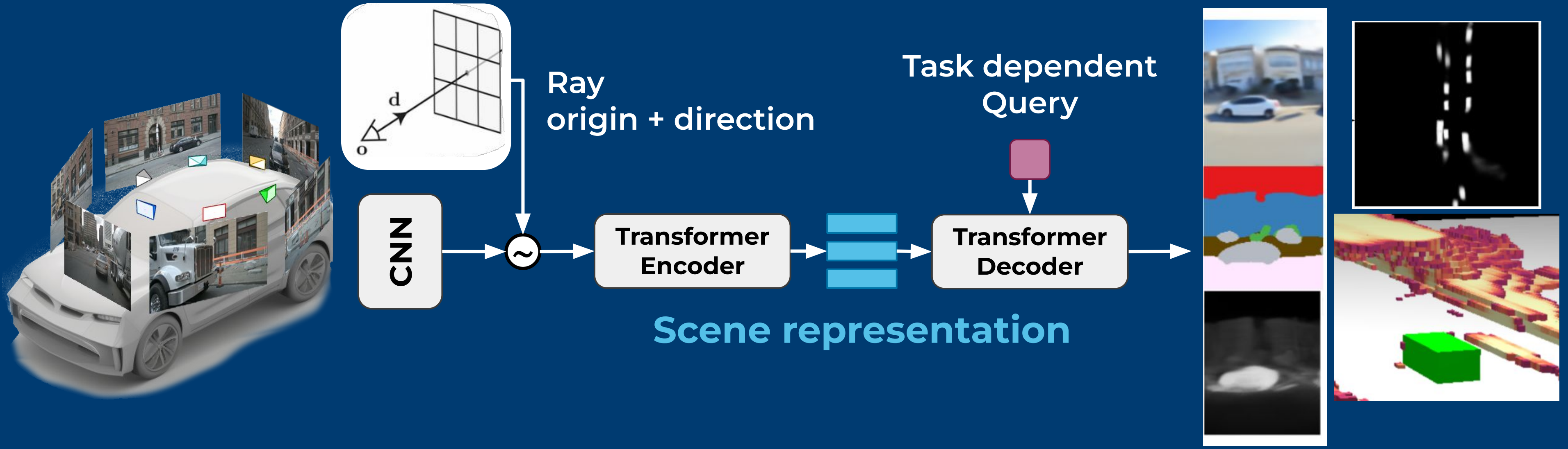






2/ What network architecture ?

Transformer offer the most flexibility

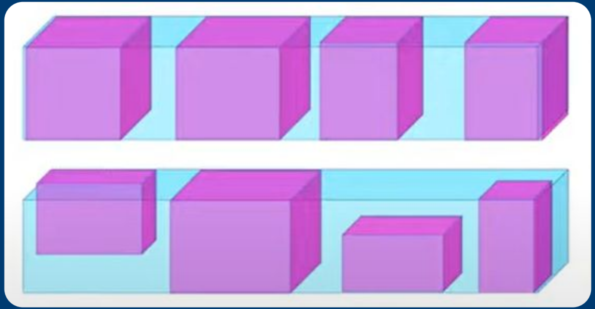


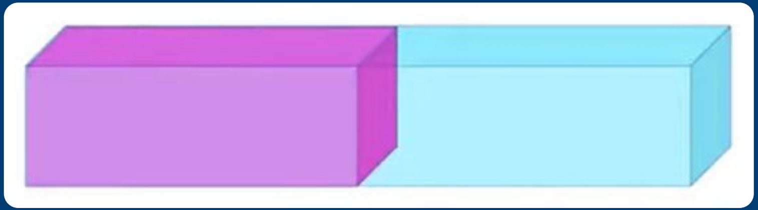
Want to learn more ?

- Bartoccioni et al., LaRa: Latents and Rays for Multi-Camera Bird's-Eye-View Semantic Segmentation, CoRL 2022 [v.ai]
- Dehghani et al., Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution
- Jaegle et al., Perceiver IO: A General Architecture for Structured Inputs & Outputs
- Sajjadi et al., Object Scene Representation Transformer

3/ What supervision ?

Self-supervision at scale





Predict masked from the visible

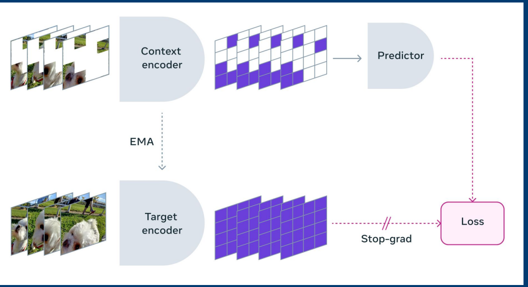
Can be used to **learn good transferable features**

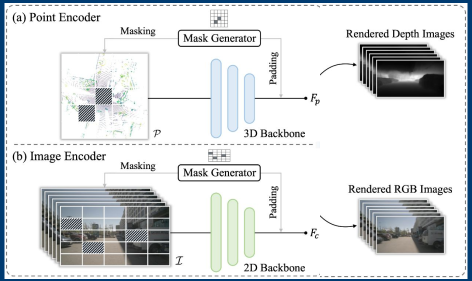
- BERT
- DINO
- V-JEPA
- UniPAD

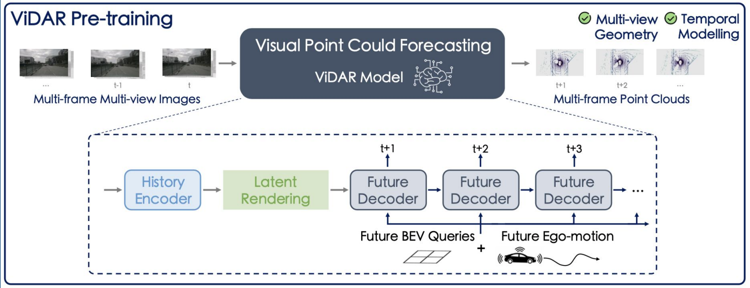
Predict future from the past

Can be used to **learn good transferable features + predictive capabilities** (forecasting, planning, control)

- GPT
- World models (e.g., GAIA-1, ViDAR, LOPR, Copilot4D)







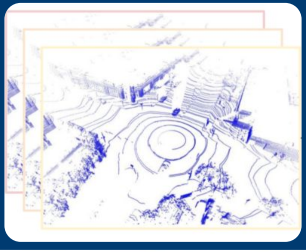
3/ What supervision ? Focus on world model

Current experimentation using principles from GPT

Language

Lorem	ipsum	dolor	sit	amet
↓	↓	↓	↓	↓
22	49	13	4	81

Sensors



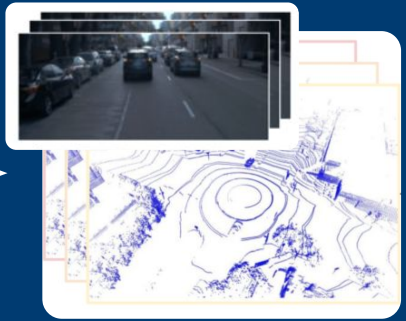
82	31	96	93	0
28	90	85	0	30
12	93	84	24	51
12	42	51	42	83
31	3	85	23	95

Past observations

→ Tokenizer → World Model → Decoder →

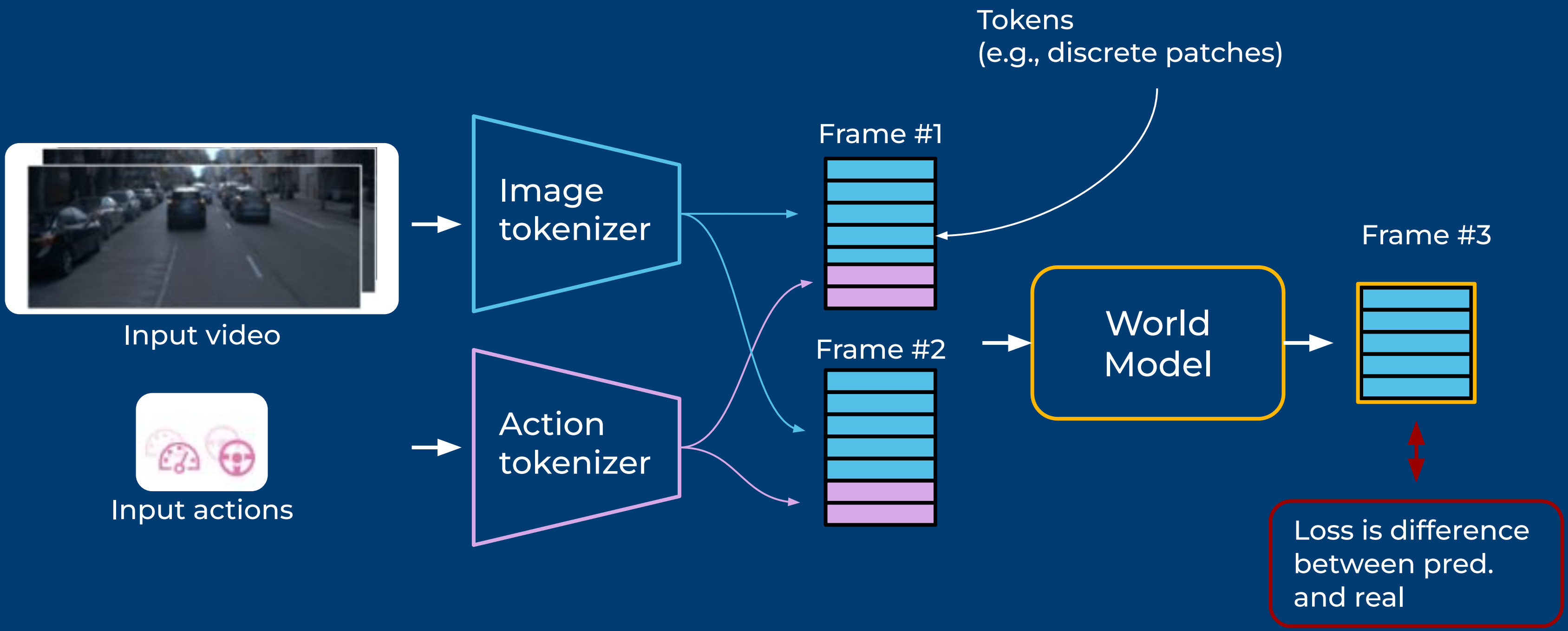
Ego action ↓ World Model

Predicted Future



3/ What supervision ? Focus on world model

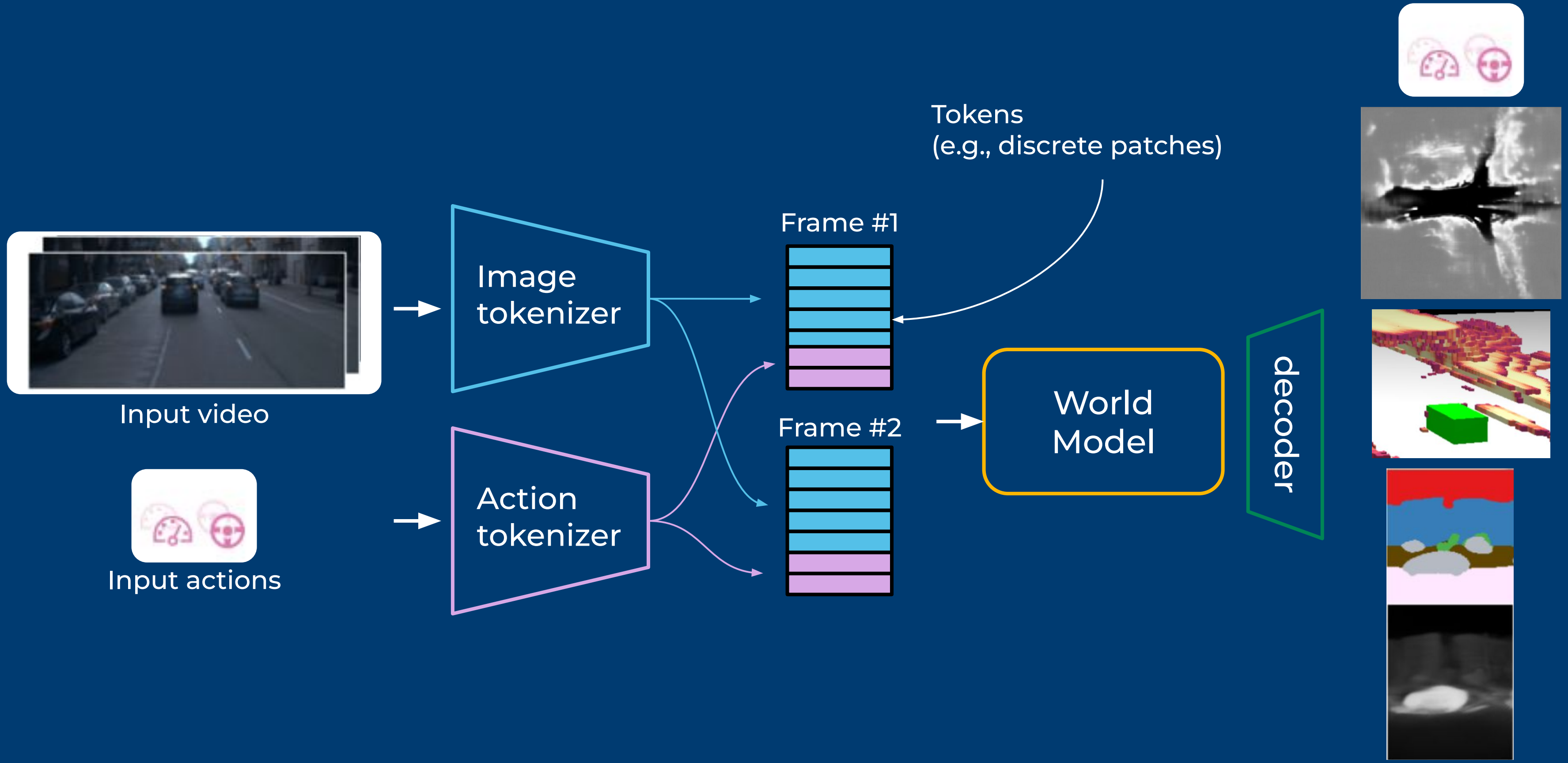
Action-conditioned predictive model



Hu et al., GAIA-1: A Generative World Model for Autonomous Driving
Zhang et al., Copilot4D: Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion

3/ What supervision ? Focus on world model

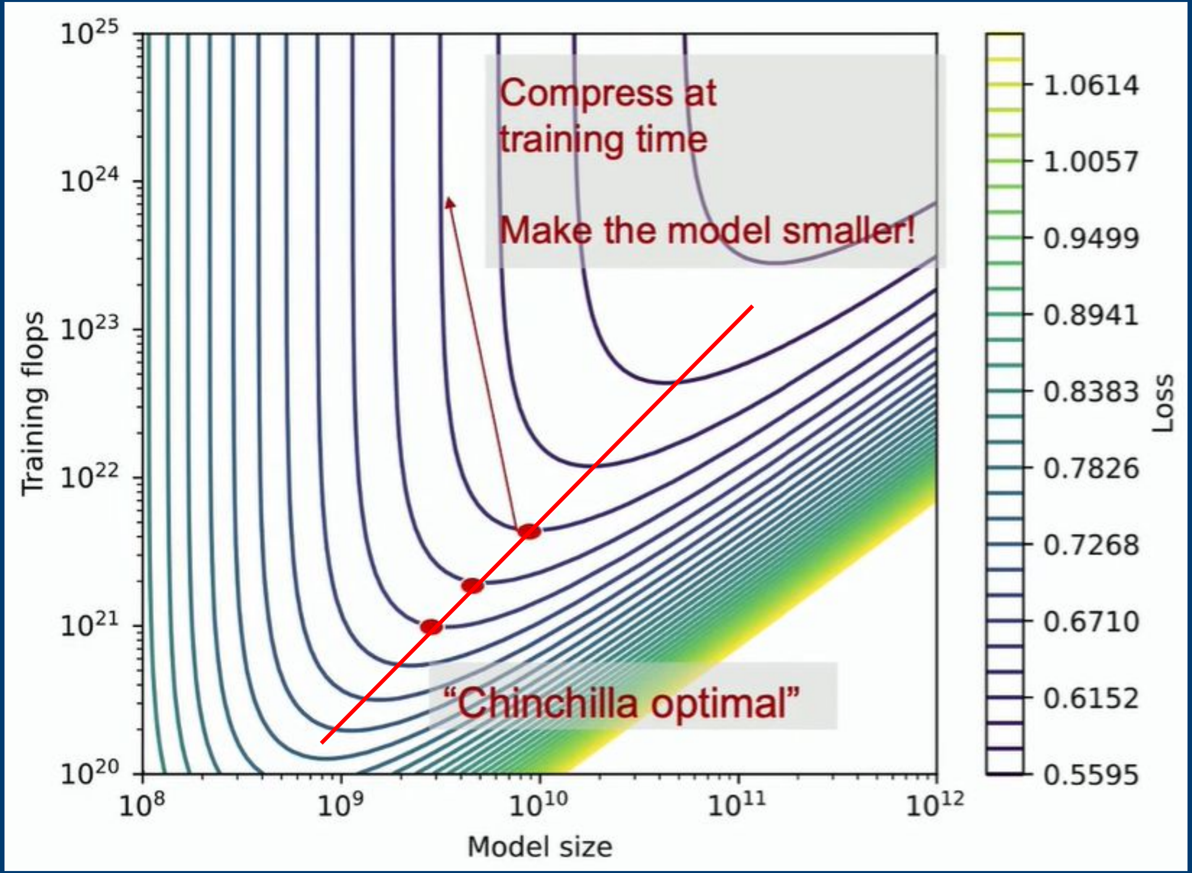
Fine-tune on any end task



4/ How to scale ? ➡ Scaling laws

Predictability in the cost/performance trade-off

Training flops ~ 6*#Params*#Training tokens



A functional approximation and a stochastic approximation

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}.$$

More weights! More tokens!

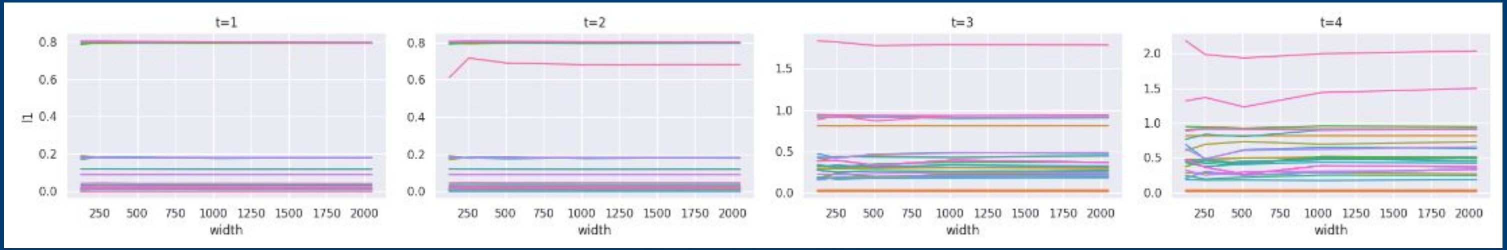
- Given a computational budget what's my best use (model size and data) ?
- At fixed TOPs how much data do I need ?

Hoffmann et al., Training Compute-Optimal Large Language Models, NeurIPS 2022

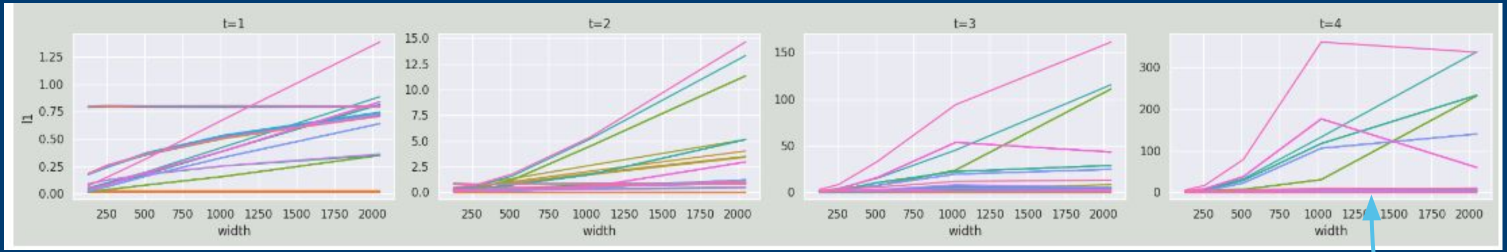
4/ How to scale ? ➡ Maximal Update Parametrization (Greg Yang et al.)

Efficient hyper parameter search + zero-shot HPs transfer

GPT2 w/ muP

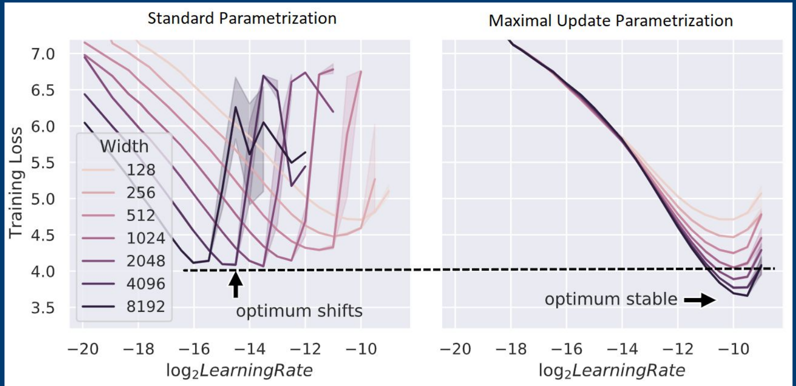


GPT2 w/o muP



tensor.abs().mean()

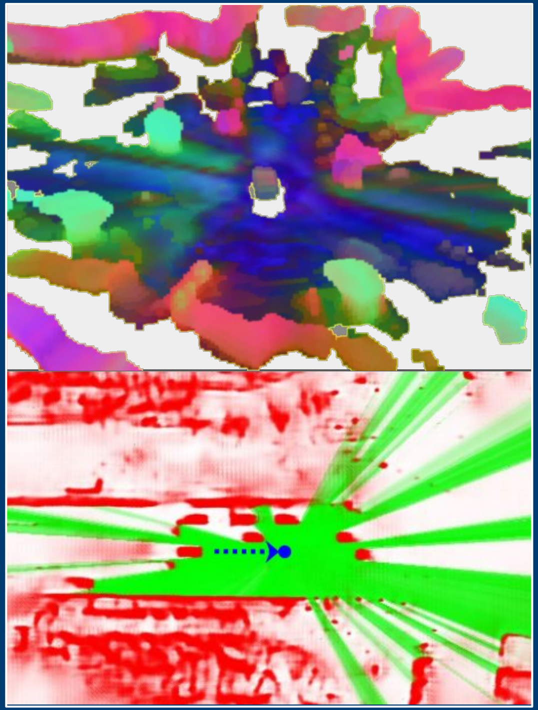
Loss vs learning rate w/ & w/o muP



Conclusion

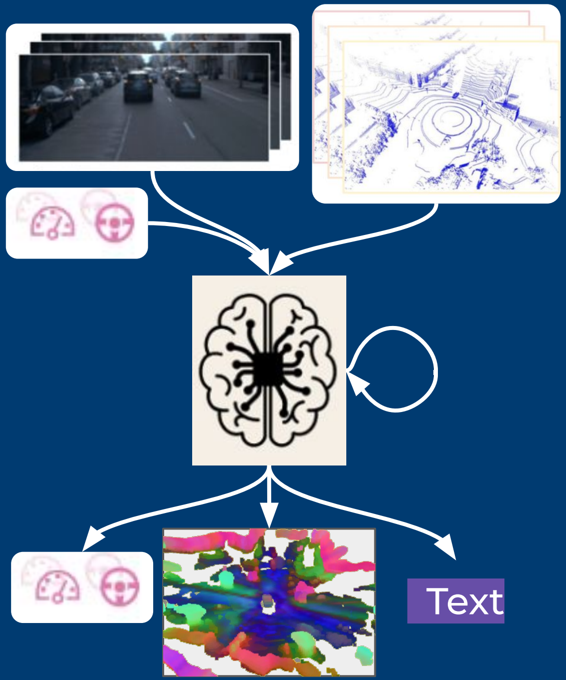
How ?

Distillation



- Explicit geometry as base
- 3D or BEV occupancy
- Features from foundation model

From Scratch

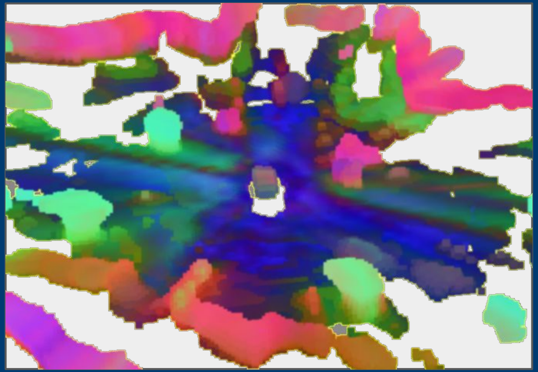


- Everything is learned
- High dimensional vectors
- Most flexible

Recap

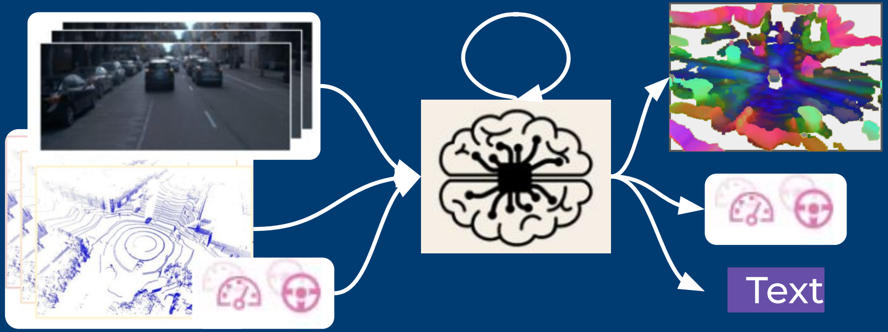
What's next ?

Distillation



- Explicit geometry as base
- 3D or BEV occupancy
- Features from foundation model
- More efficient
- Less control

World Model



- Everything is learned
- More control, More flexible
- Higher cost

Final notes

- Not possible without JZ and Adastra
- World models are a promising avenue for robotics
- Need to study more their behaviors
- Extension to multimodality and generalisation to different rigs