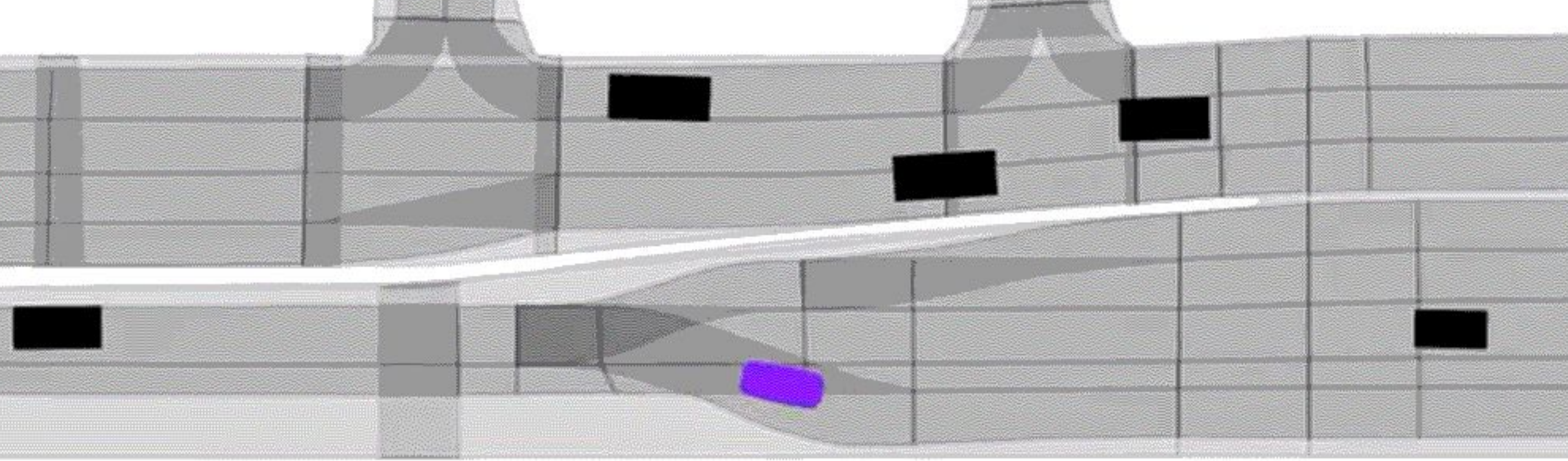


Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion

ICLR 2024

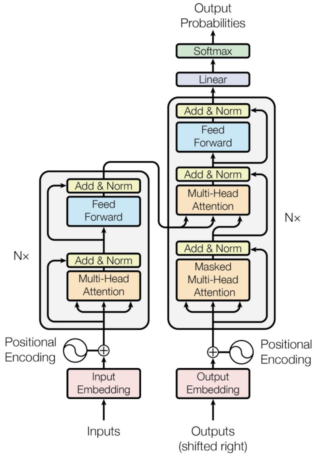
Lunjun Zhang   Yuwen Xiong   Ze Yang   Sergio Casas   Rui Hu
Raquel Urtasun

UNIVERSITY OF TORONTO

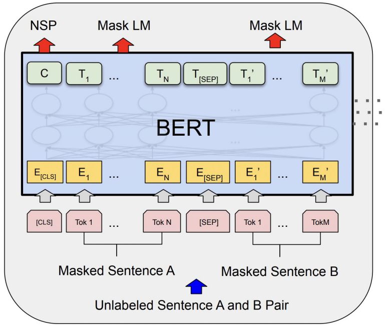
waabi

The Era of Foundation Models

NLP:



Transformer
(Vaswani et al, 2017)



Pre-training

BERT
(Devlin et al, 2018)

Language Models are Few-Shot Learners

GPT-3
(Brown et al, 2020)

Vision:

BigGAN
(Brock et al, 2018)

DALL-E 2
(Ramesh et al, 2022)

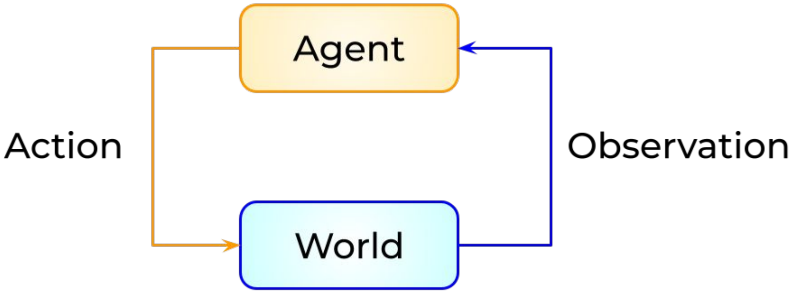
Segment Anything
(Kirillov et al, 2023)

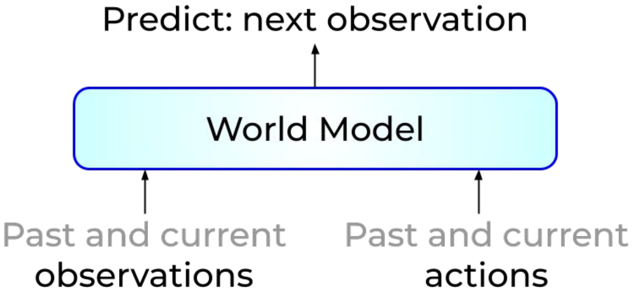
Robotics:

What foundation model should robotics scale?

Learning a World Model

- World Models **predict the next observation** in an environment given the current action and the past observations.
- Learning a world model is an **unsupervised learning** process: it requires no labels or rewards.
- This idea has been around for a long time, dating back to adaptive control and model-based reinforcement learning.





waabi

Bottlenecks of Scaling World Models

- Training World Models to predict the next observation is very similar to training **Language Models** to **predict the next token.**
- What **bottlenecks** held us back from scaling unsupervised world models on robotic applications such as autonomous driving?
  - Why hasn't it become the *default* model to train for robotics?

Predicting in **complex** and **unstructured** observation space

The **scalability** of the generative model

A Scalable Recipe for Learning World Models

Two bottlenecks:

Predicting in **complex** and **unstructured** observation space

The **scalability** of the generative model

Solution:

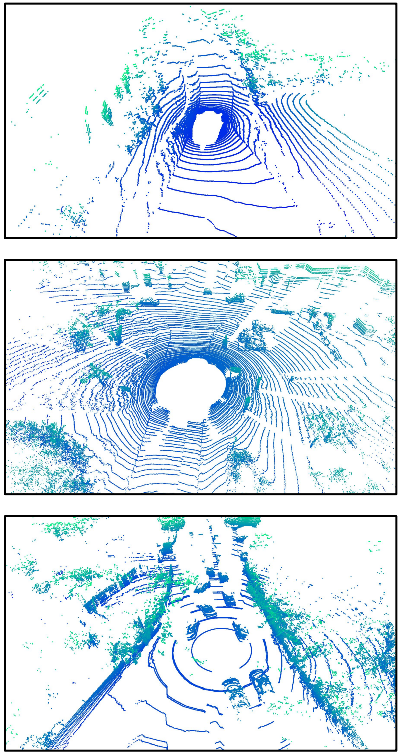
Tokenize Everything

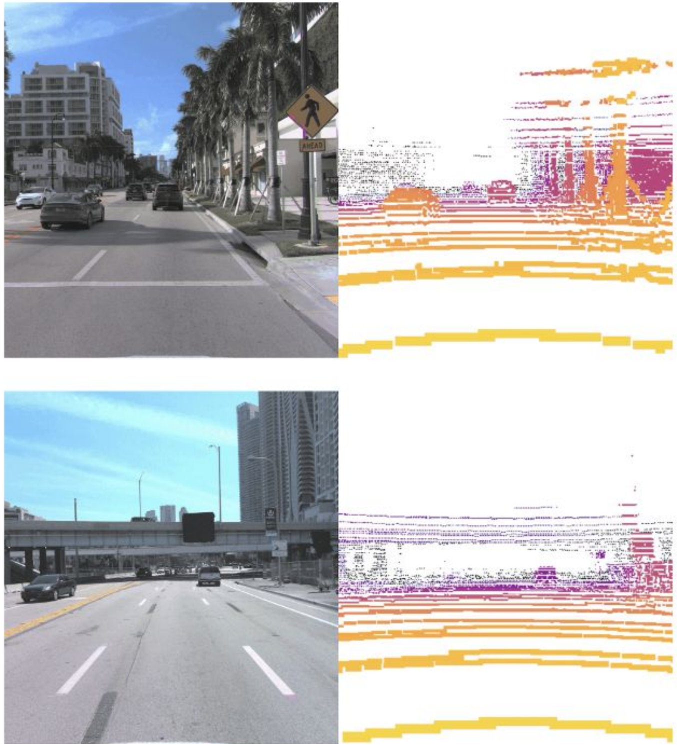
Discrete Diffusion

waabi

Bottleneck 1: Complex / Unstructured Observation Space

Designing a generative model that captures **meaningful likelihoods** can be highly non-trivial!





Self-Driving Datasets

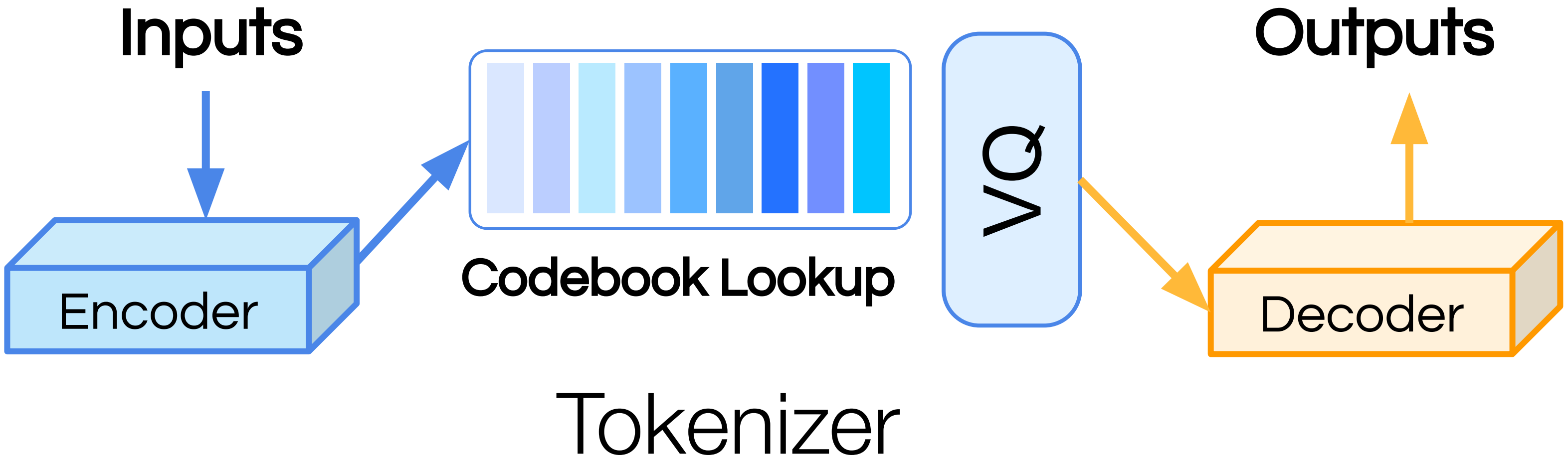
KITTI (Geiger et al, 2013);
NuScenes (Caesar et al, 2019);
Argoverse 2 (Wilson et al, 2023)

Solution: Tokenize Everything

Designing a generative model that captures **meaningful likelihoods** can be highly non-trivial!

By contrast, **language models** first **tokenize** a text corpus, then predict **discrete** indices like a classifier.

Our Solution: train a **VQVAE** to **tokenize everything**.

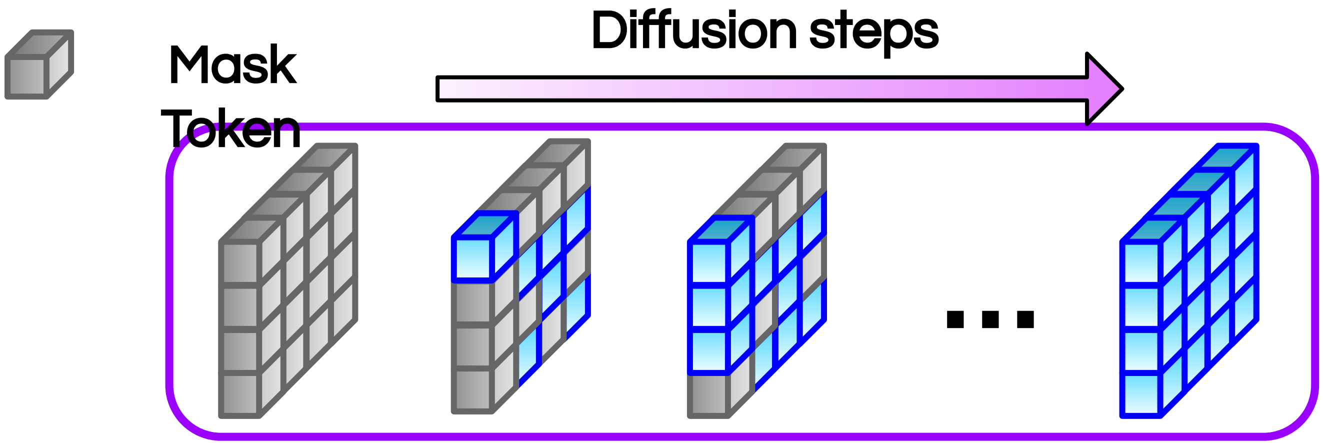


Bottleneck 2: Scalability of the Generative Model

- Autoregressive GPT training can be applied on any tokenized data, but with one problem: GPTs only decode one token at a time.
- In robotics, **a single observation has tens of thousands of tokens**, so parallel decoding of tokens becomes a must.
    - Decoding all the tokens of an observation in parallel would **incorrectly** assume that all those tokens are conditionally independent given past observations.

Solution: Discrete Diffusion

- Discrete diffusion is a natural solution to this problem.
  - Decodes **arbitrary** number of tokens at each step
  - Can **iteratively refine** the already decoded tokens



Austin et al, “Structured Denoising Diffusion Models in Discrete State-Spaces”, 2021

Chang et al, “MaskGIT: Masked Generative Image Transformer”, 2022.

waabi

Discrete Diffusion Made Simple

- We modify the popular Masked Generative Image Transformer (MaskGIT) into an **absorbing-uniform discrete diffusion** model.
- It is essentially a BERT trained to **both infill and denoise**.

Algorithm 1 Training

repeat
  $\mathbf{x}_0 : \{1, \cdots, |V|\}^N \sim q(\mathbf{x}_0)$
  $u_0 \sim \text{Uniform}(0, 1)$
  Randomly mask $\lceil \gamma(u_0) N \rceil$ tokens in $\mathbf{x}_0$
  $u_1 \sim \text{Uniform}(0, 1)$
  Randomly noise $(u_1 \cdot \eta)\%$ of remaining tokens
  $\mathbf{x}_k \leftarrow$ *masked-and-noised* $\mathbf{x}_0$
  $\arg\max_\theta \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_k)$ with cross entropy
until converged

Algorithm 2 Sampling

$\mathbf{x}_K =$ all mask tokens
for $k = K - 1, \ldots, 0$ **do**
  $\tilde{\mathbf{x}}_0 \sim p_\theta(\cdot \mid \mathbf{x}_{k+1})$
  $\boldsymbol{l}_k = \log p_\theta(\tilde{\mathbf{x}}_0 \mid \mathbf{x}_{k+1}) + Gumbel(0, 1) \cdot k/K$
  On non-mask indices of $\mathbf{x}_{k+1}$: $\boldsymbol{l}_k \leftarrow +\infty$
  $M = \lceil \gamma(k/K) N \rceil$
  $\mathbf{x}_k \leftarrow \tilde{\mathbf{x}}_0$ on top-M indices of $\boldsymbol{l}_k$
end for
return $\mathbf{x}_0$

Chang et al, "MaskGIT: Masked Generative Image Transformer", 2022.

Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018.

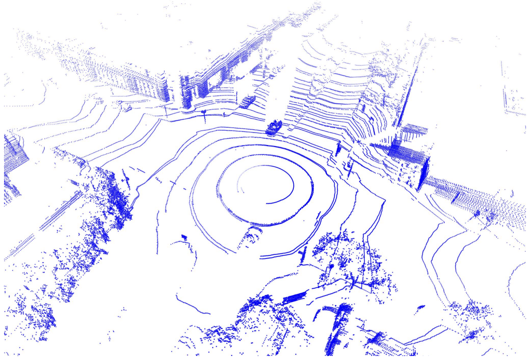
What foundation model should robotics scale?

Our proposal for learning an **unsupervised world model**:

- Tokenize everything by training VQVAE
- Discrete diffusion as the core generative model
- Learn to predict the future

Tokenize the 3D World for Autonomous Driving

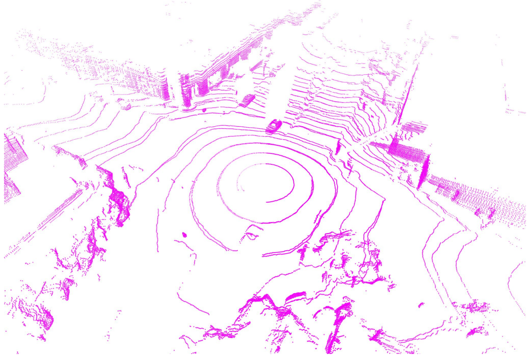
Observation

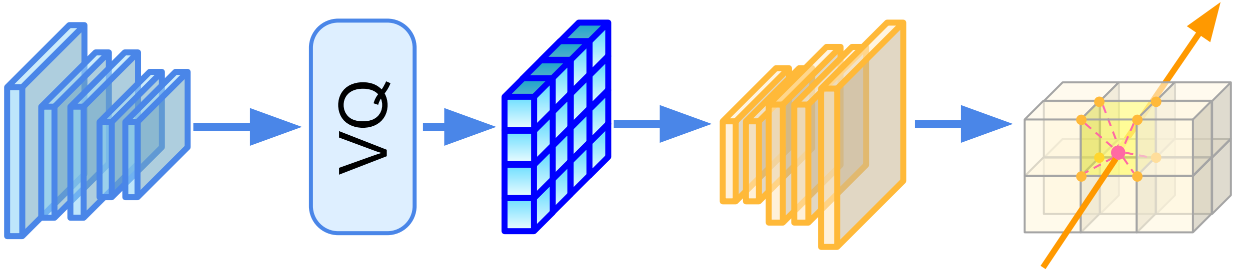


Tokenize
the 3D World

VQVAE

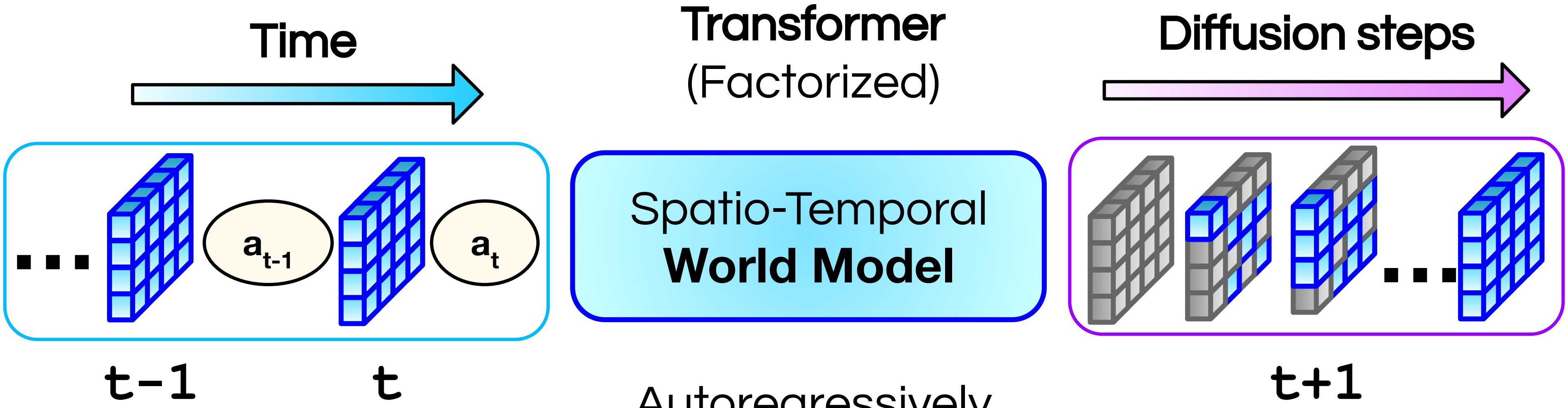
Reconstruction





Bird-Eye View (BEV) Tokens

Unsupervised 4D World Model for Autonomous Driving



Autoregressively
predict future frames

waabi

A Mixture of Training Objectives

We train the world model on a **mixture** of training objectives

- 50% of the time: condition on the past, predict the future.
- 40% of the time, denoise the past and the future jointly.
- 10% of the time, denoise each frame individually.

The last one enables **classifier-free diffusion guidance** at inference.

Results

- When applied to learning world models on point cloud observations, our model **reduces prior SOTA Chamfer distance by more than 65% for 1s prediction, and more than 50% for 3s prediction.**

NuScenes

NuScenes 1s	Chamfer↓	L1 Med↓	AbsRel Med↓
SPFNet	2.24	-	-
S2Net	1.70	-	-
4D-Occ	1.41	0.26	4.02
Ours	**0.36**	**0.10**	**1.30**
NuScenes 3s			
SPFNet	2.50	-	-
S2Net	2.06	-	-
4D-Occ	1.40	0.43	6.88
Ours	**0.58**	**0.14**	**1.86**

KITTI

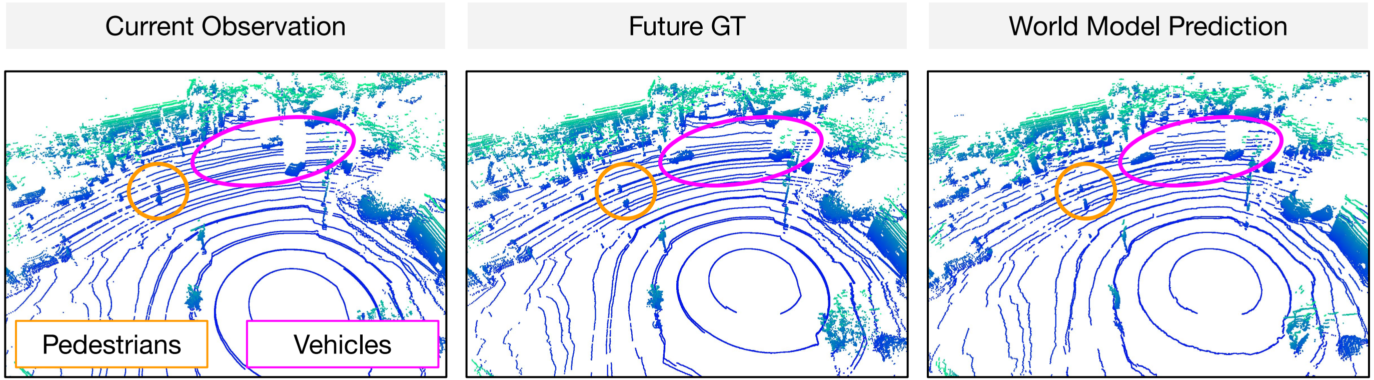
KITTI 1s	Chamfer↓	L1 Med↓	AbsRel Med↓
ST3DCNN	4.11	-	-
4D-Occ	0.51	0.20	2.52
Ours	**0.18**	**0.11**	**1.32**
KITTI 3s			
ST3DCNN	4.19	-	-
4D-Occ	0.96	0.32	3.99
Ours	**0.45**	**0.17**	**2.18**

Argoverse 2

1s Prediction	Chamfer↓	L1 Med↓	AbsRel Med↓
4D-Occ	1.42	0.24	1.67
Ours	**0.26**	**0.15**	**0.94**
3s Prediction			
4D-Occ	1.99	0.42	2.88
Ours	**0.55**	**0.19**	**1.26**

Visualizations

- Highly accurate Accurate Near-Term 1s Prediction

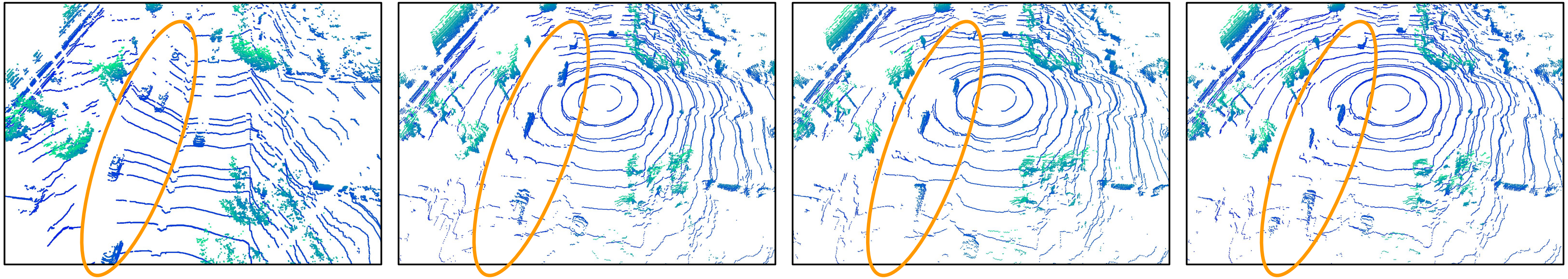


waabi

Visualizations

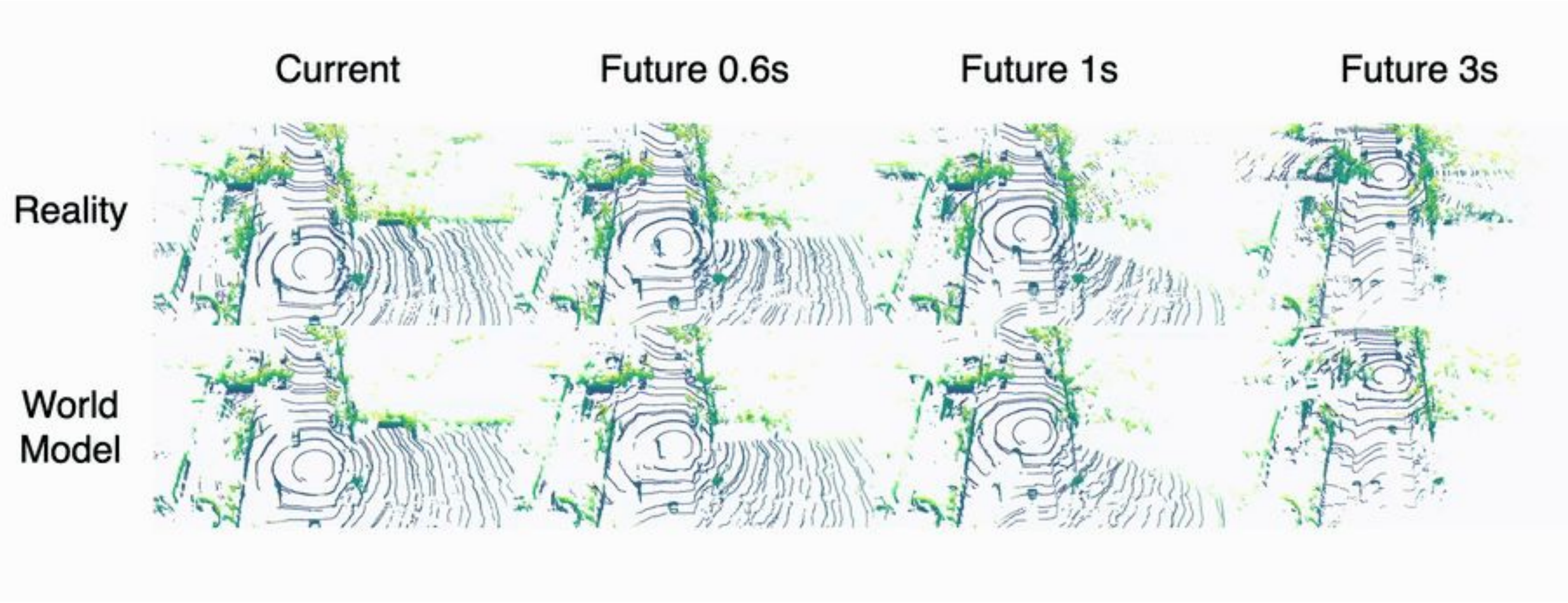
- Diverse Multi-Future 3s Prediction

Current Observation	World Model Future 1	World Model Future 2	World Model Future 3

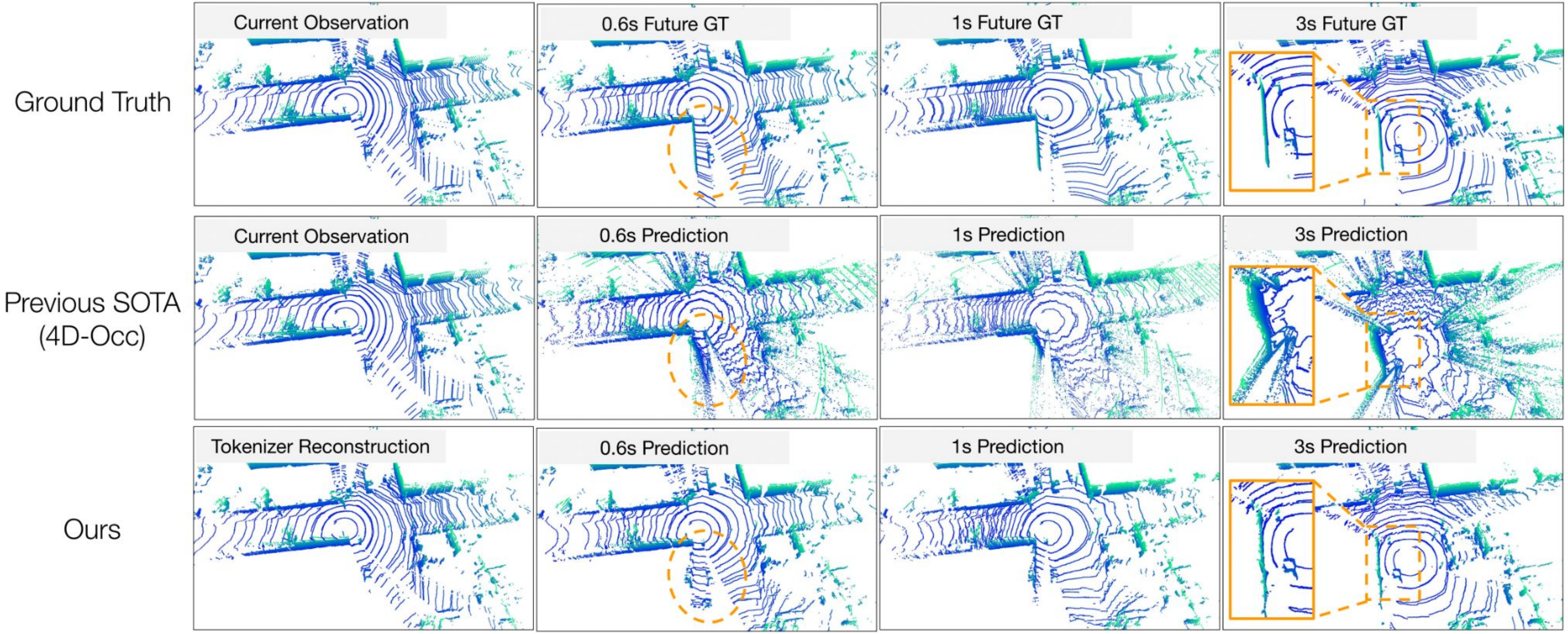


Traffic on the other side of road

Visualizations

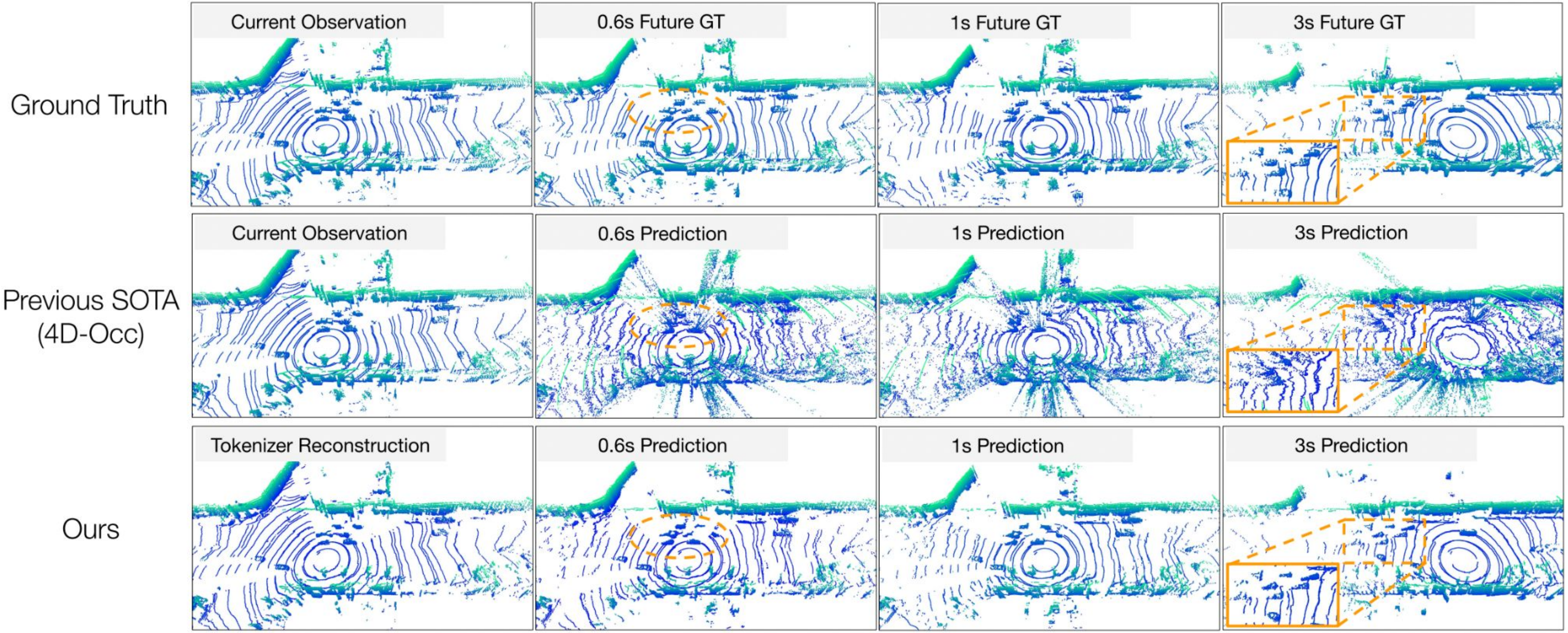


Qualitative Comparisons



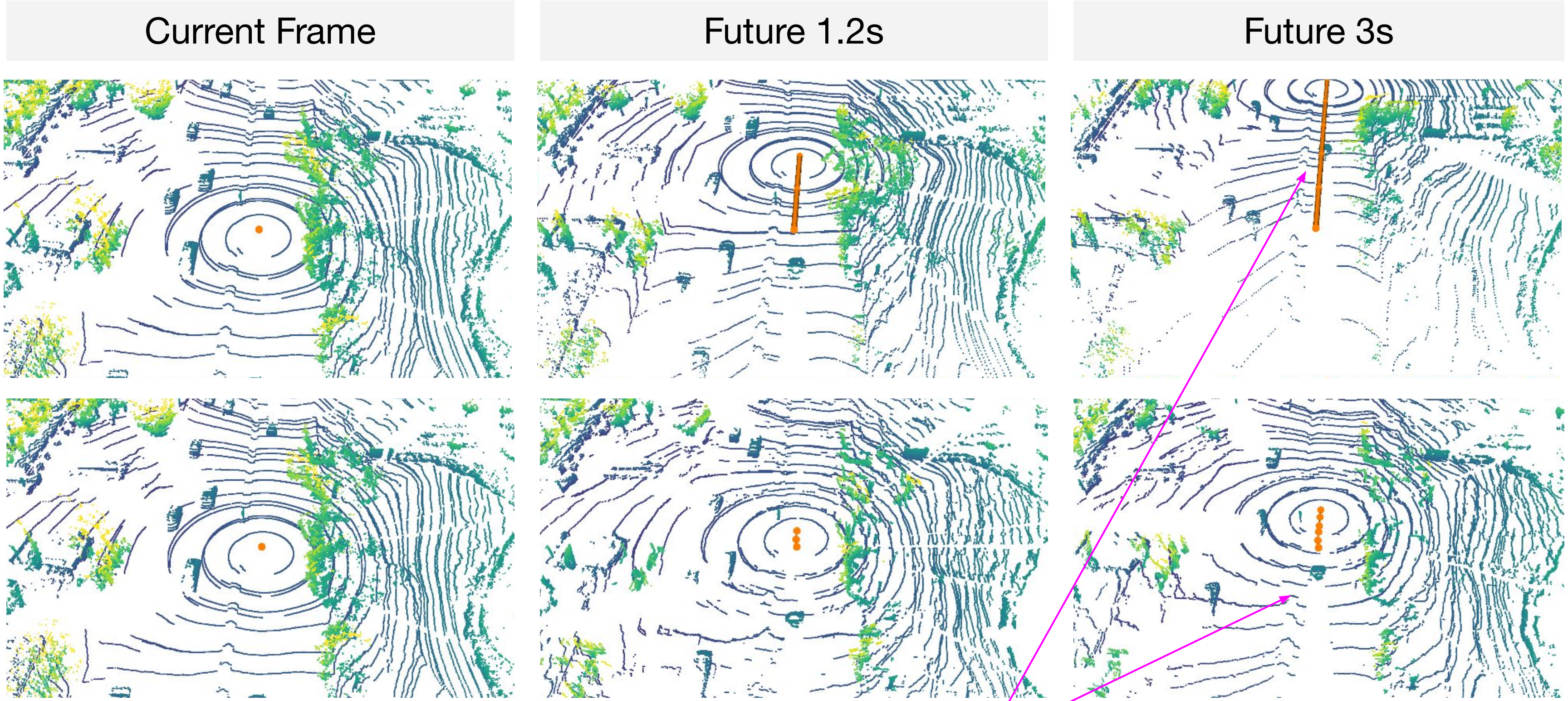
Khurana et al, “Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting”, 2023.

Qualitative Comparisons



Khurana et al, "Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting", 2023.

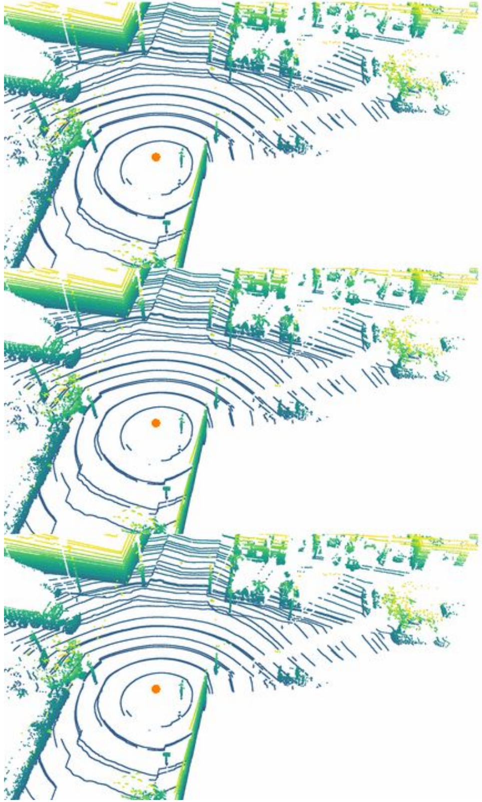
Evaluating Counterfactual Actions



Counterfactual action: the ego vehicle brakes.
World model prediction**: the vehicle behind will also brake to avoid collision.**

Evaluating Counterfactual Actions

Current



Ground Truth

Prediction

Counterfactual

Conclusion

- Learning unsupervised world models is a promising way to build foundation models for robotics
- We propose a highly effective recipe for learning world models: **Tokenize Everything** + Discrete Diffusion + Spatio-Temporal Transformer
- When applied to the point cloud forecasting task in autonomous driving, our method achieves SOTA results
- Remains an open question on how such a world model can directly improve the decision making capabilities of robotic agents

waabi