Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion

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

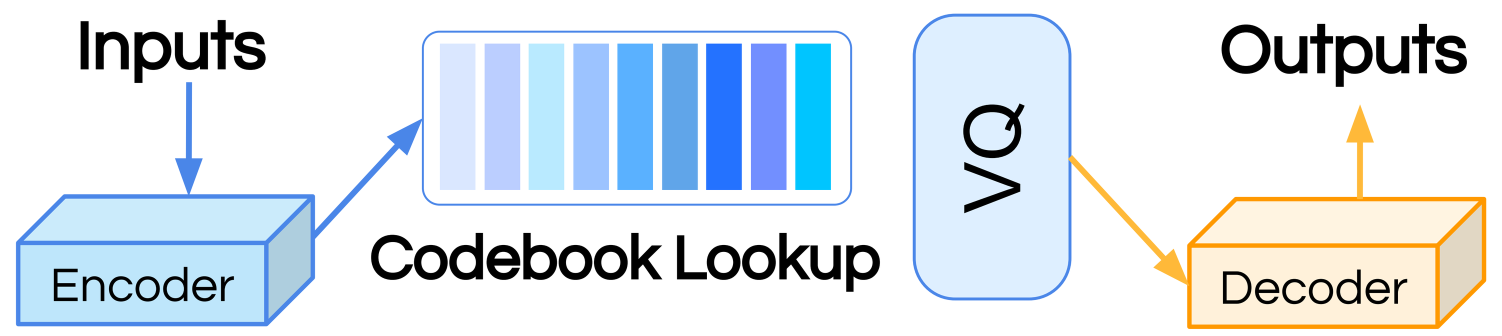
waabi · University of Toronto

What Model Should Robotics Scale?

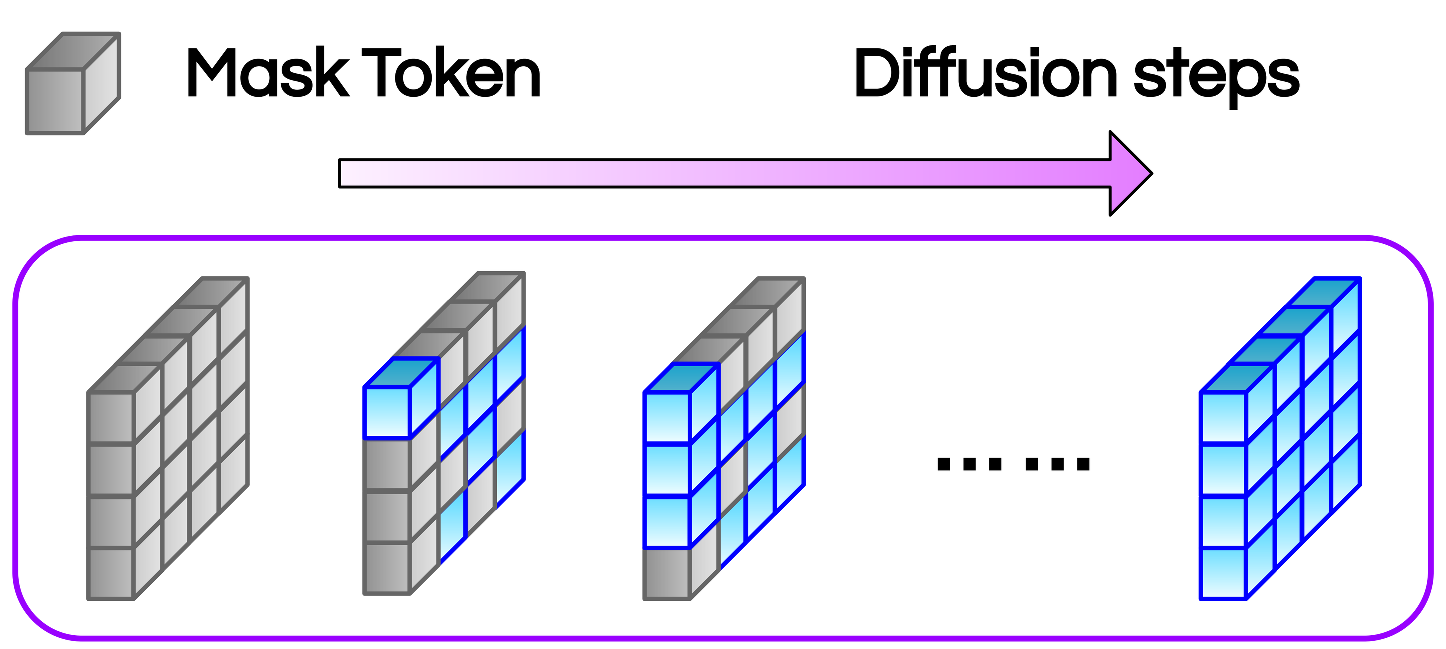
- NLP scales **GPT** by simply predicting the next token.
- Similarly, we can train **a robotics foundation model** to learn how the world works by predicting the future.
- What bottlenecks have been previously holding us back from scaling **unsupervised world models** on robotic applications such as autonomous driving?

Bottlenecks of Scaling World Models

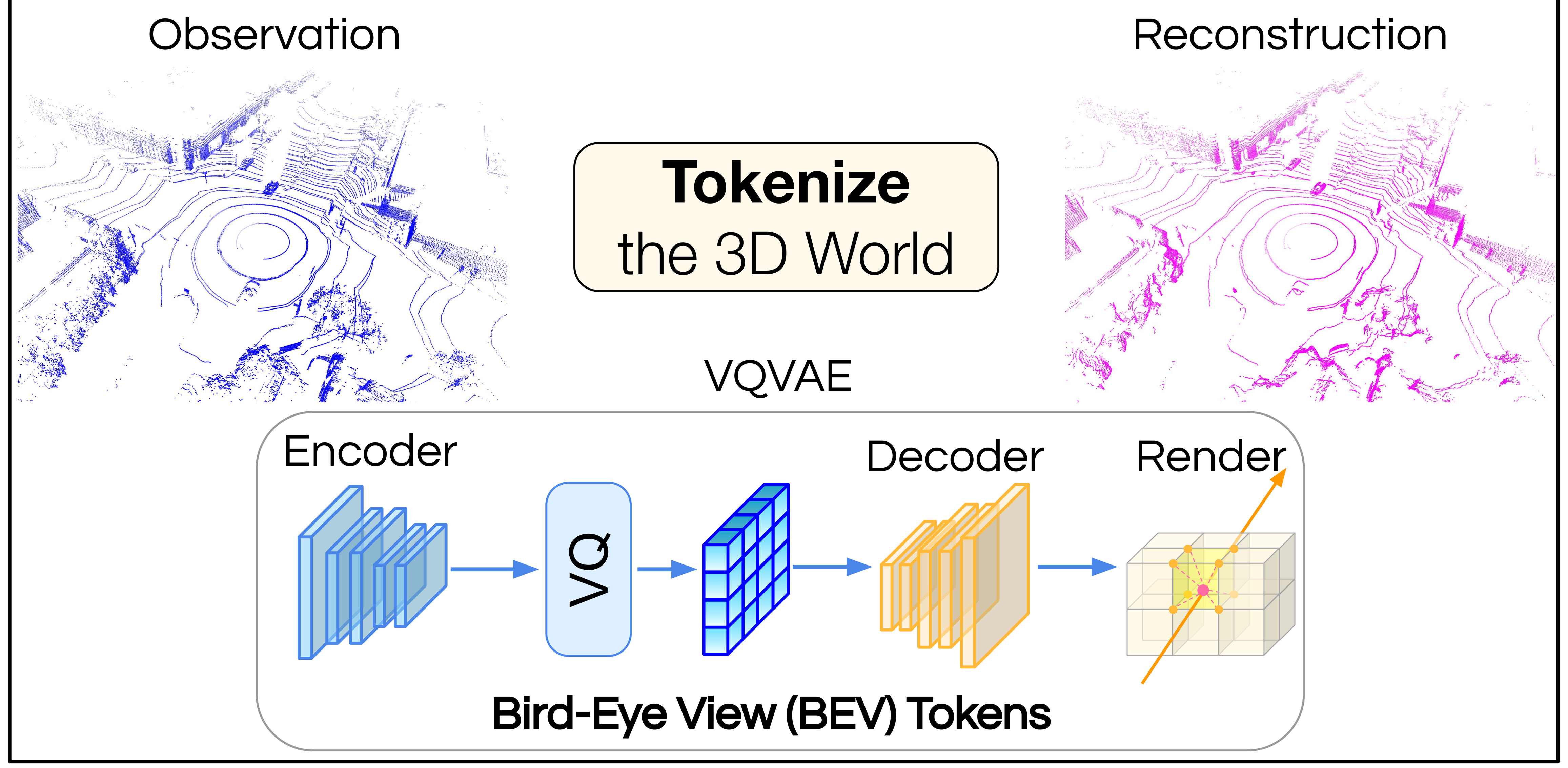
- **Bottleneck 1**: complex, unstructured observation space.
  - By contrast, language models first tokenize a text corpus, then predict discrete indices like a classifier.
  - **Our Solution: train a VQVAE to tokenize everything**.



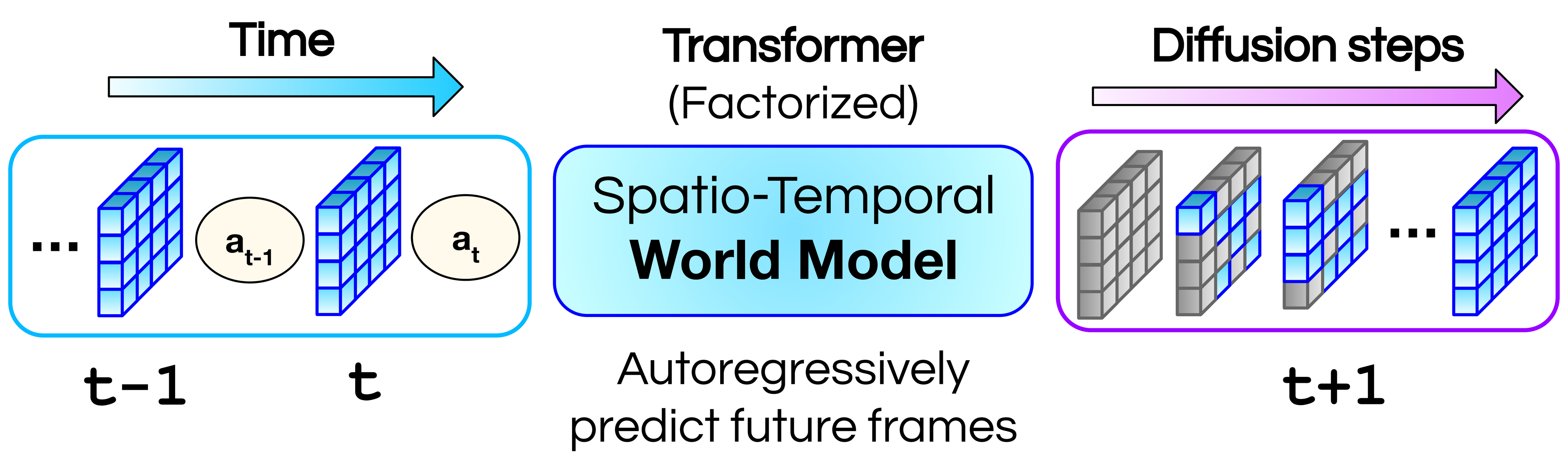
- **Bottleneck 2**: scalability of generative models.
  - Language models are known to scale well, but they only decode one token at a time..
  - In autonomous driving, **a single observation has tens of thousands of tokens**.
  - **Our Solution: discrete diffusion** rather than autoregressive modeling.



Tokenize the 3D World



Unsupervised 4D World Model for Autonomous Driving



Discrete Diffusion Made Simple

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** $\mathbf{x}_0$

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.**

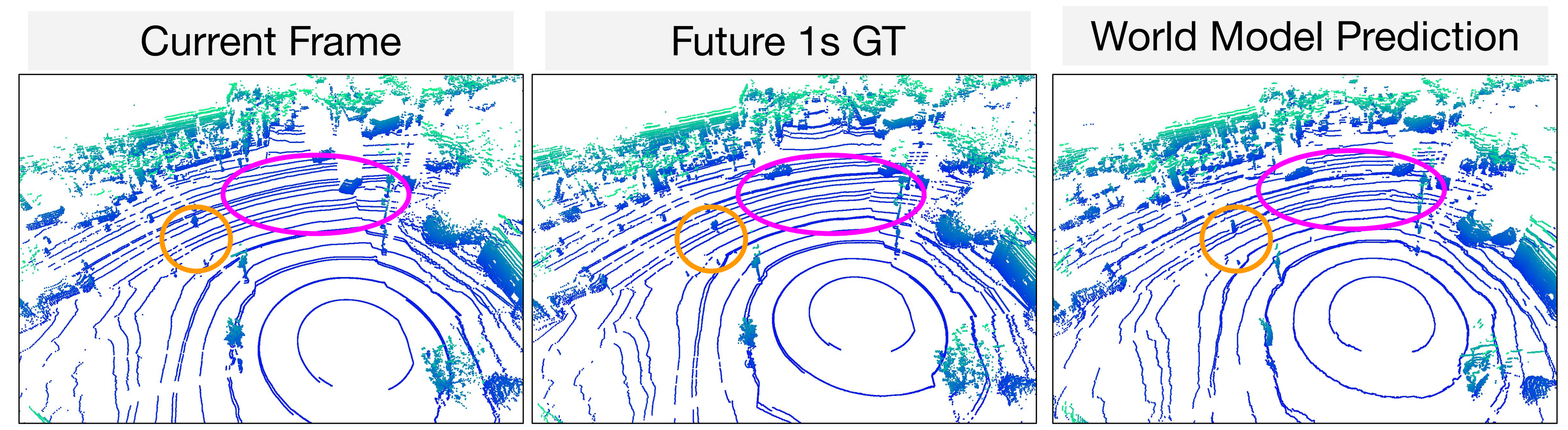
NuScenes & KITTI

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 1s			
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**

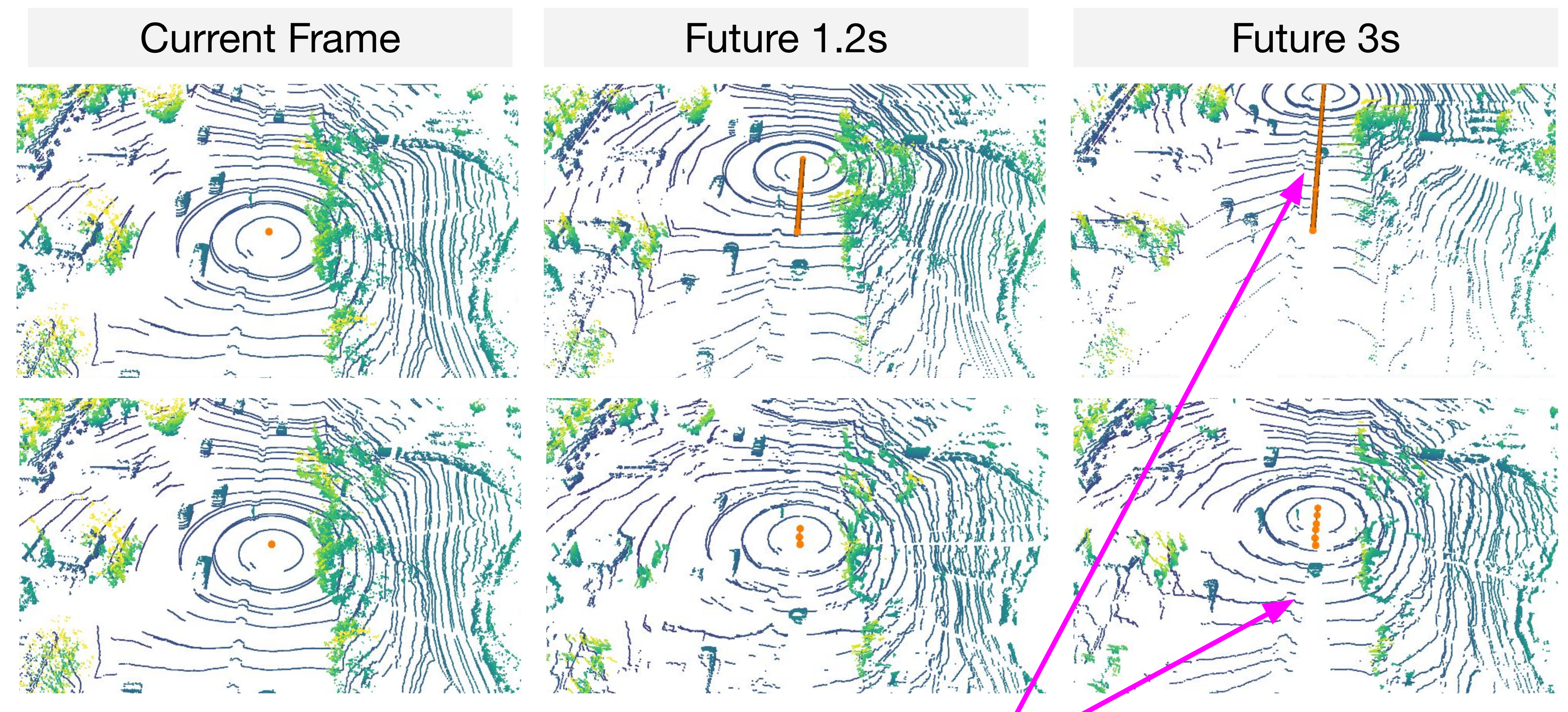
Argoverse 2

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**

Tricks that matter: 1. Classifier-free diffusion guidance, using the entire past history as the "prompt"; **2.** Train to predict an entire segment of the future (not just 1 frame).



Evaluating Counterfactual Actions



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