Masked Generative Priors Improve World Models Sequence Modelling Capabilities

ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling, 2025

Cristian Meo*
Mircea Tudor Lică*
Zarif Ikram*
Akihiro Nakano
Vedant Shah
Aniket Rajiv Didolkar
Dianbo Liu
Anirudh Goyal
Justin Dauwels

Outstanding Paper Award!

World models

- Deep RL has achieved breakthrough performance in complex tasks.
- World models enable sample efficiency via “imagination”

Basic Training Formula…

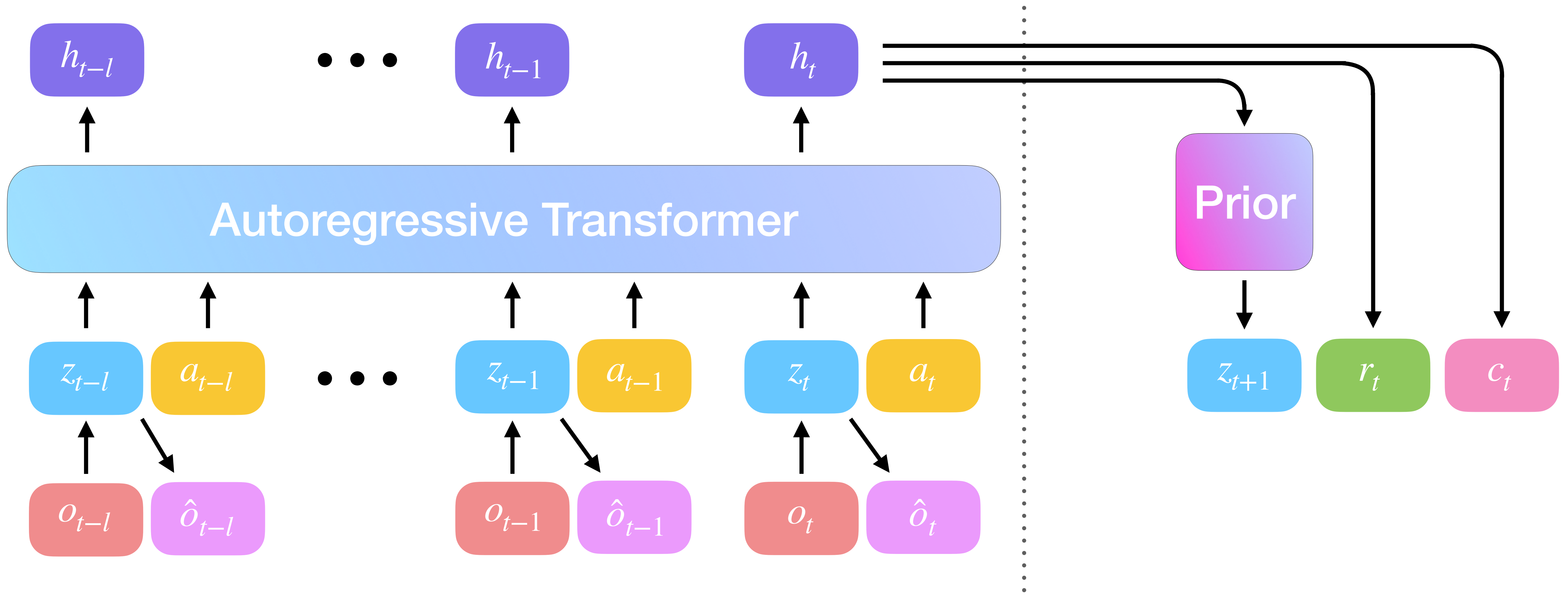
For visual observations

- Pixel space encoder
    - Turns observations to latent embeddings
- Sequence modeler
    - $f : \mathscr{S} \times \mathscr{A} \rightarrow \mathscr{S}$
- Learnt end-to-end through
    - Some reconstruction such as observation, latent, etc.

Autoregressive Transformer

As a sequence modeller

- Unidirectional generation process
  - Unable to fully capture global contexts
- e.g., STORM[1],
  - $h_t \rightarrow z_{t+1}$ is a MLP
  - Predicting categorical logits



Zhang, Weipu, et al. "Storm: Efficient stochastic transformer based world models for reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2023): 27147-27166.

What could go wrong?

- No global information
- Only *one* chance to predict z_t
  - What if $logit_1$ and $logit_2$ need to be distinct?
- Result=Predicting infeasible states!

Masked Generative Modelling

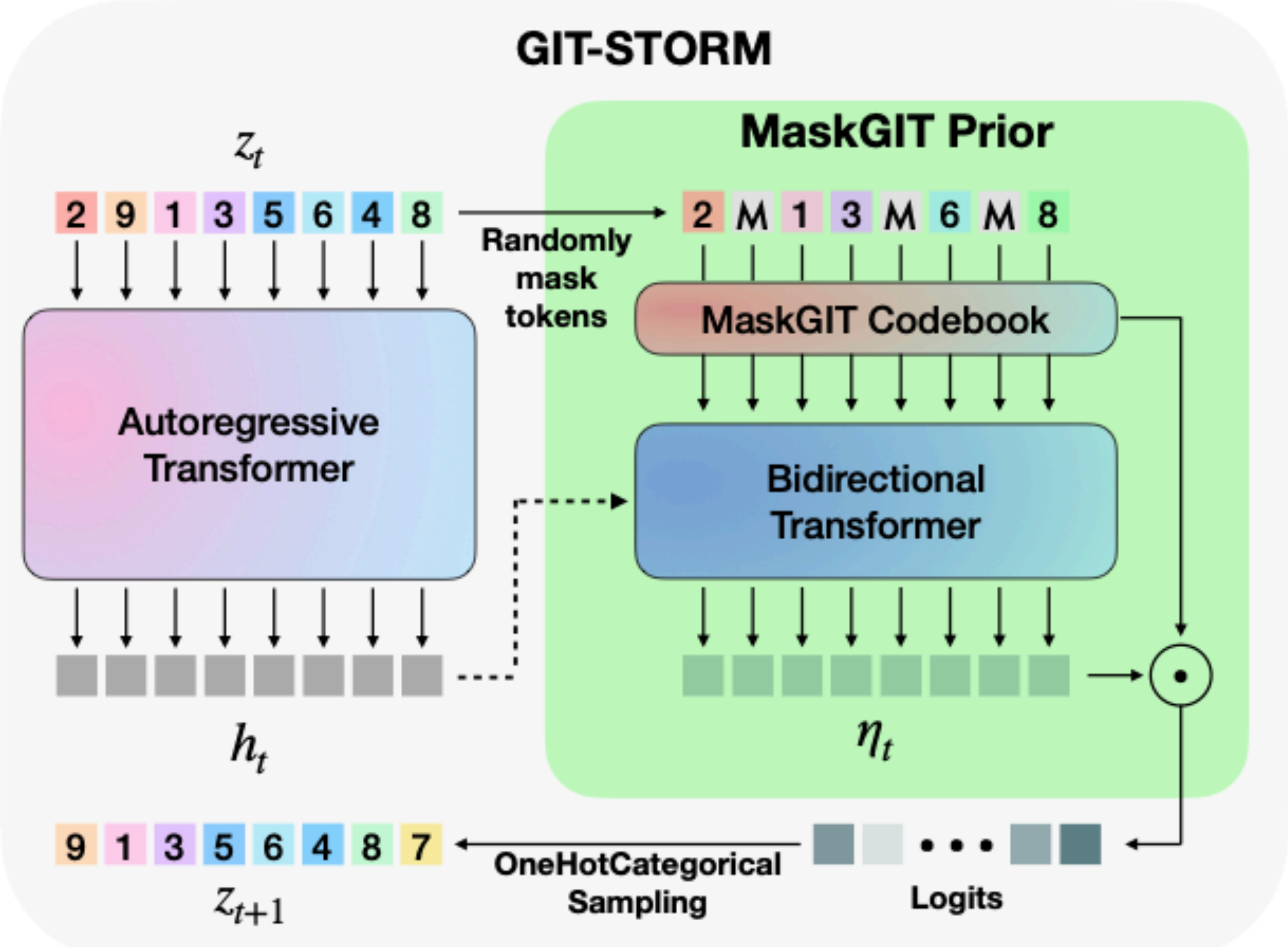
- TECO [2] introduces MaskGIT[3] prior $p_\phi(z_{t+1} \mid h_t)$
    - draft-and-revise predicts the next discrete representations
- Uses global context
    - "Safer" approach to predict z_t
    - One step at a time (after looking at current global prediction)

Yan, Wilson, et al. "Temporally consistent transformers for video generation." International Conference on Machine Learning. PMLR, 2023.
Chang, Huiwen, et al. "Maskgit: Masked generative image transformer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

World model w/ masked generative prior

tl;dr: we replace a MLP module with a MaskGIT module

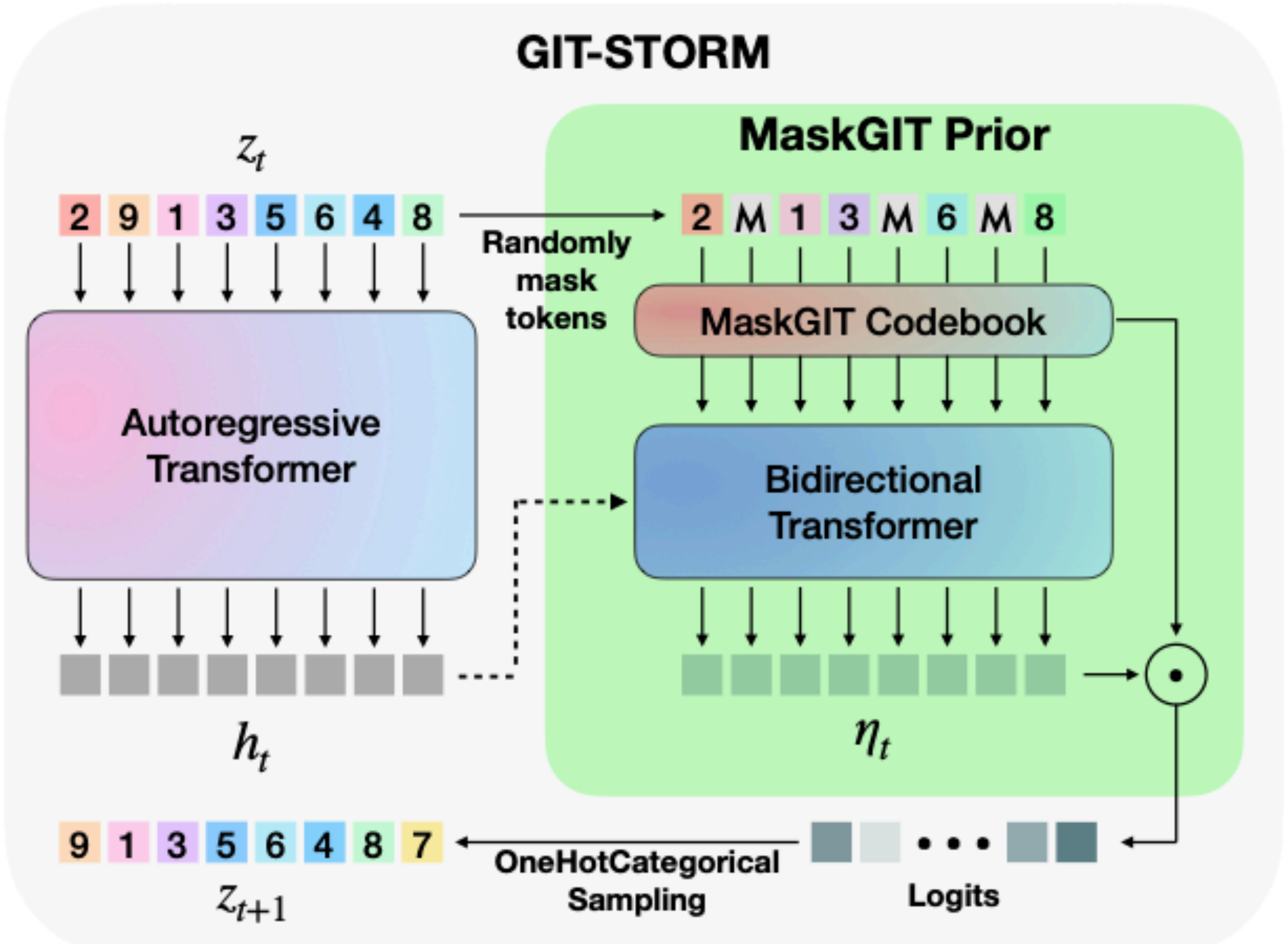
- Concatenate hidden state h_t with masked latent representation $m_t \cdot z_t$
  - Posterior latent prediction from masked latent z_t
- Minimize KL div. Between prior and post.
  - Similar to previous methods



World model w/ masked generative prior

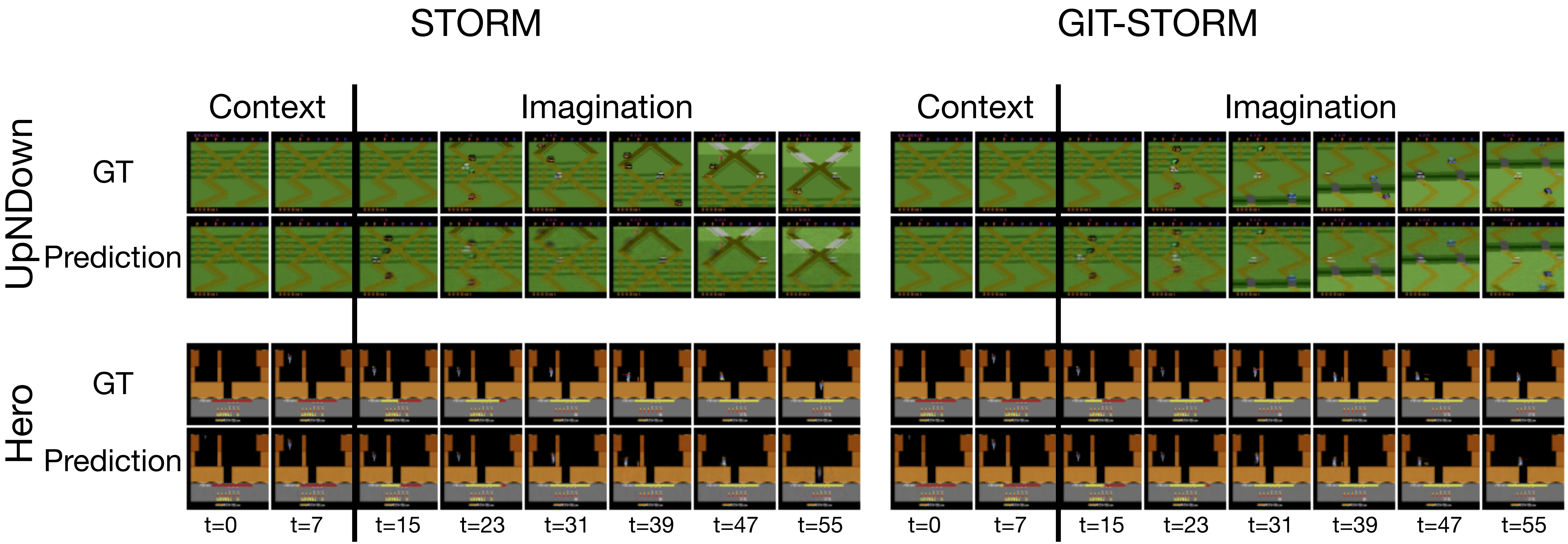
tl;dr: we replace a MLP module with a MaskGIT module

- During inference
    - perform draft-and-revise using masked decoding
- Don't generate everything at once
    - or do next-token-prediction
- Mask tokens, predict the next ones, and revise as needed



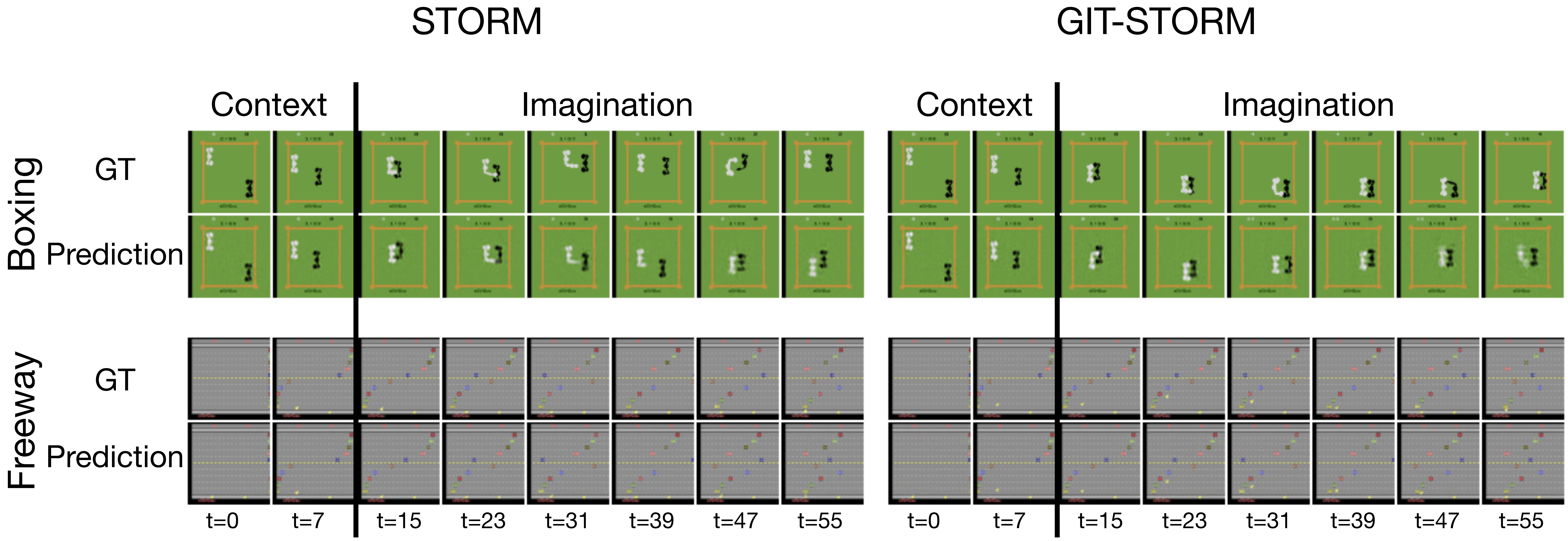
Results

Better modeling capabilities!



Results

Better modeling capabilities!



Results

Better modeling capabilities!

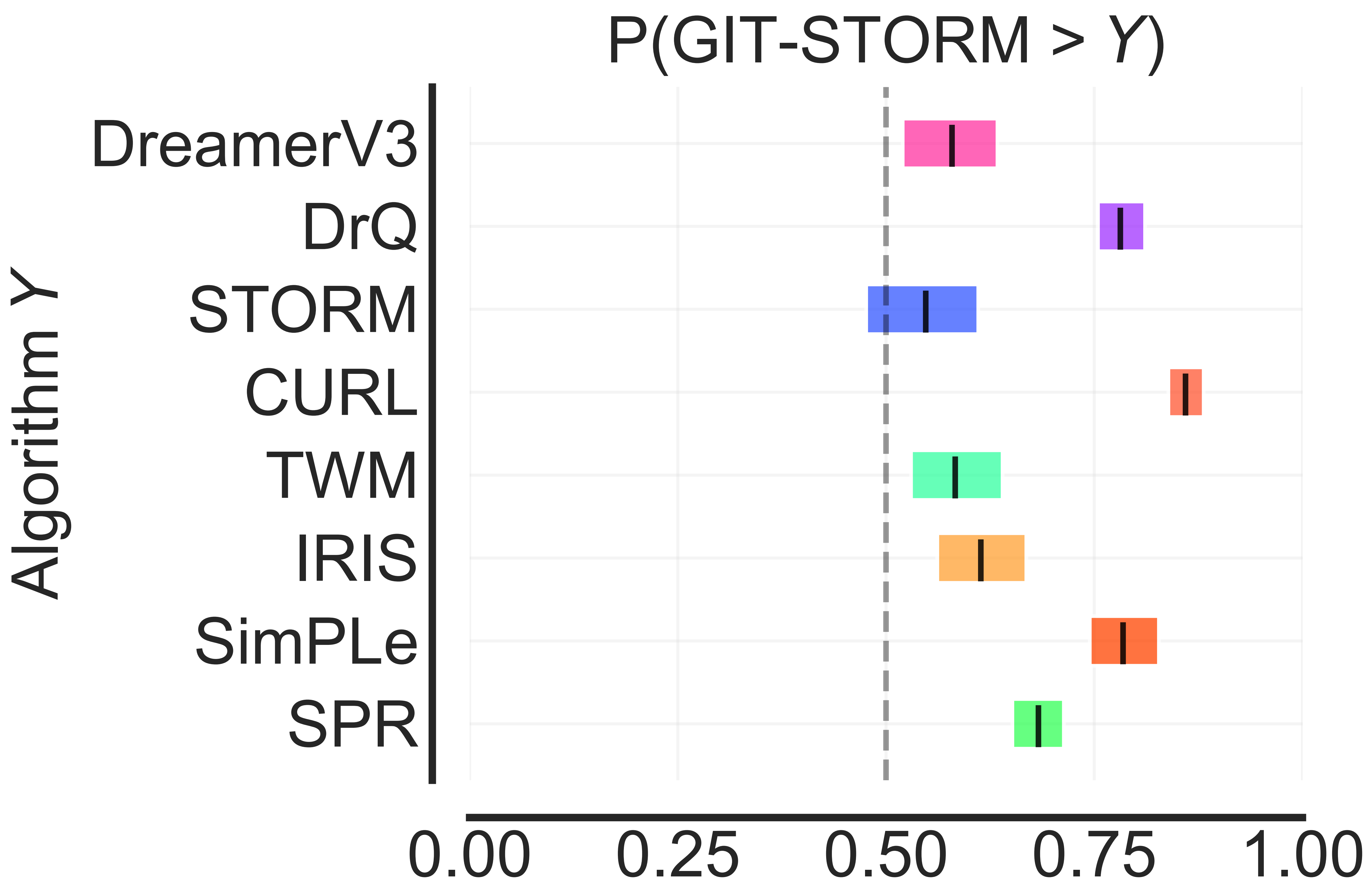
Game	FVD (↓)		Perplexity (↑)	
	STORM	GIT-STORM	STORM	GIT-STORM
Boxing	**1458.32**	1580.32	49.24	**54.95**
Hero	381.16	**354.16**	10.55	**30.25**
Freeway	105.45	**80.33**	33.15	**67.92**

Task	FVD (↓)		Perplexity (↑)	
	STORM	GIT-STORM	STORM	GIT-STORM
Cartpole Balance Sparse	2924.81	**1892.44**	1.00	**3.76**
Hopper Hop	4024.11	**3458.19**	3.39	**22.59**
Quadruped Run	3560.33	**1000.91**	1.00	**2.61**

Results

… Leading to better policy

- **19%** higher human mean than STORM
- Over **20%** improved IQM than STORM
- Over **50%** probability of improvement to all baselines



Results

Works for continuous action environments too

- Transformer-based world models to continuous action spaces (DMC Suite) was unaddressed by IRIS, TECO or the original STORM

- GIT-STORM reports results on DMC benchmark

    - Over **50%** probability of improvement to STORM, PPO, and SAC

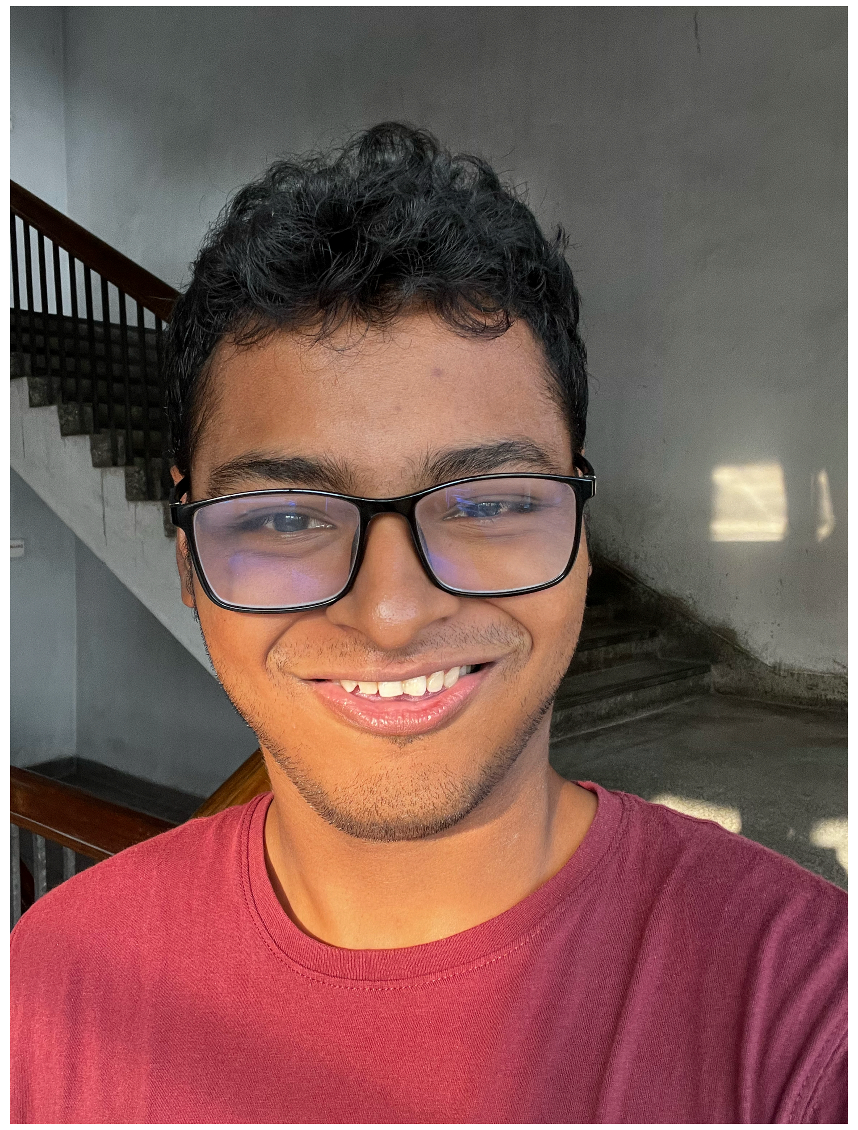
Limitations and Future Works

- Our method falls short on continuous action space benchmarks, compared to GRU based approaches
  - *Why does transformer based methods fails to capture continuous action space worlds?*
- We *can* use *only* one iteration for the Draft-and-Revise decoding scheme
  - *How to fully exploit the advantages of this decoding scheme?*

Summary

- World modeling approaches involve categorical distribution for latents
- Treat prior distributions as “2d grid”
  - Just like images
- Learn to uncover “mass” for the distributions using maskGIT
  - Refine prior distributions during inference

Thank You!



Zarif Ikram
Undergraduate Visiting Scholar, NUS
Incoming PhD Student, USC Viterbi

Cristian Meo
PhD Student, TU Delft

Paper

Repository