Video Generation Models as World Simulators

January 8th, 2025

Junoh Kang

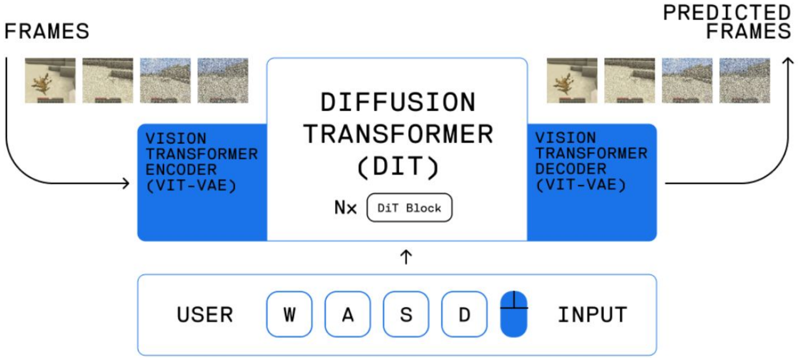
World Simulator?

What is world simulator

A **world simulator** is a system or model that can generate, replicate, or predict aspects of a physical or conceptual environment, enabling users to explore, interact with, or analyze it as if it were the real world. These simulators aim to mimic the dynamics, rules, and interactions of complex systems, making them useful for training, research, testing, and creative purposes.

World Simulator?

OASIS: A Universe in a Transformer



- Playable
- Realtime
- Open-world AI model

OASIS: A Universe in a Transformer

What are conditions for video generation models to function as playable world simulator or game engine?

- **Long generation**
- **Low latency**
- …

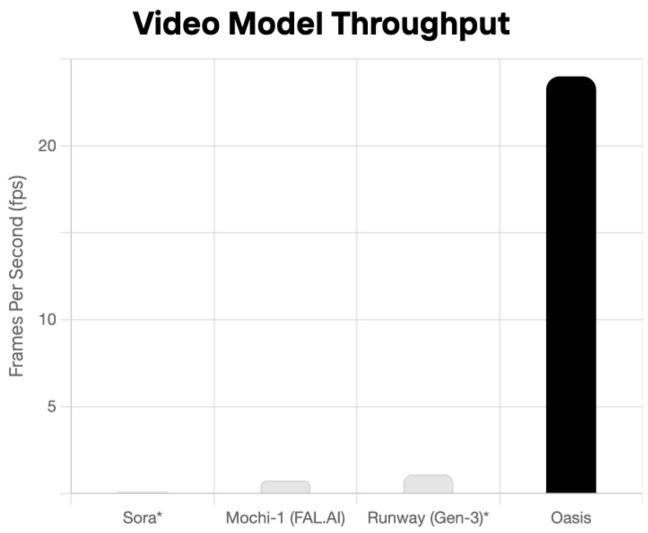
To achieve these, OASIS trains models following Diffusion-Forcing.

Their presentation at NeurIPS : https://neurips.cc/Expo/Conferences/2024/talk%20panel/100361

OASIS: A Universe in a Transformer

Low Latency

- Optimized kernel for hardware



Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion

Boyuan Chen
MIT CSAIL
boyuanc@mit.edu

Diego Marti Monso*
Technical University of Munich
diego.marti@tum.de

Yilun Du
MIT CSAIL
yilundu@mit.edu

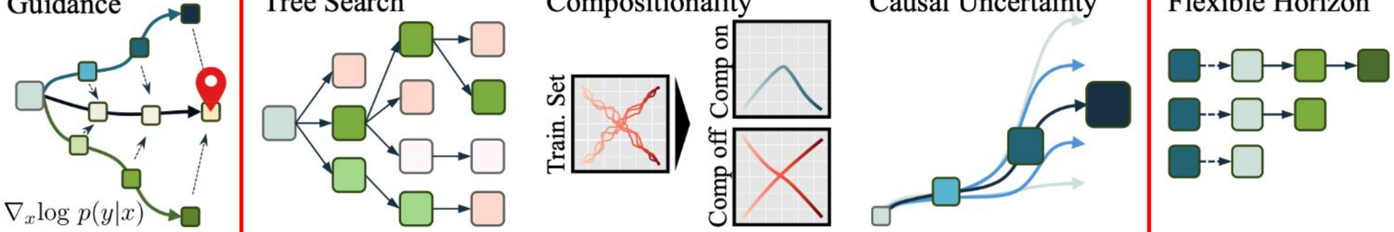
Max Simchowitz
MIT CSAIL
msimchow@mit.edu

Russ Tedrake
MIT CSAIL
russt@mit.edu

Vincent Sitzmann
MIT CSAIL
sitzmann@mit.edu

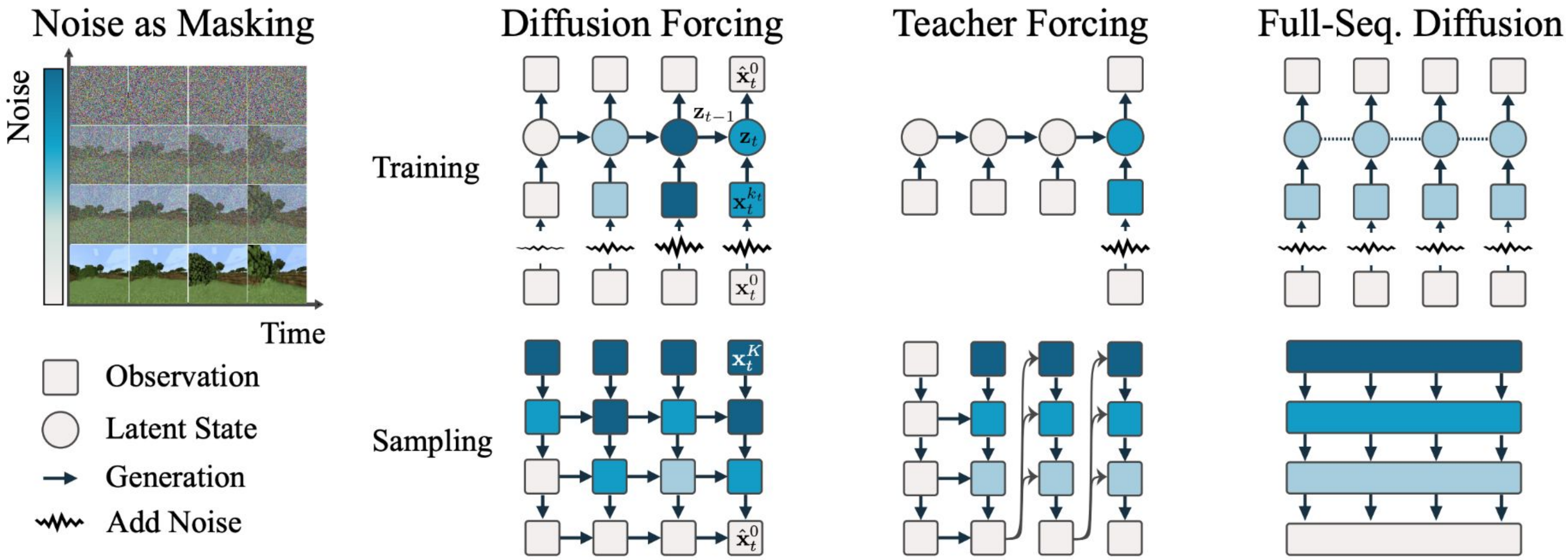
NeurIPS 2024

Diffusion Forcing



	Guidance	Tree Search	Compositionality	Causal Uncertainty	Flexible Horizon
Teacher Forcing	✗	✓	✓	✓	✓
Full-Seq. Diffusion	✓	✗	✗	✗	✗
Diffusion Forcing	✓	✓	✓	✓	✓

Diffusion Forcing = Teacher Forcing + Diffusion Models



Teacher Forcing

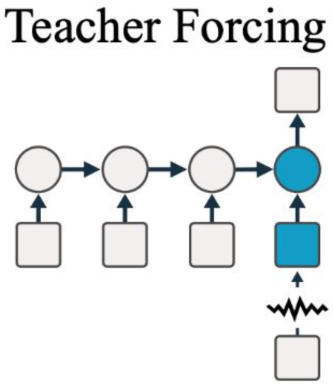
Models predict the **next token** based on a **ground truth history** of previous tokens.

(+) Flexible time horizon

(-) Unstable on continuous data

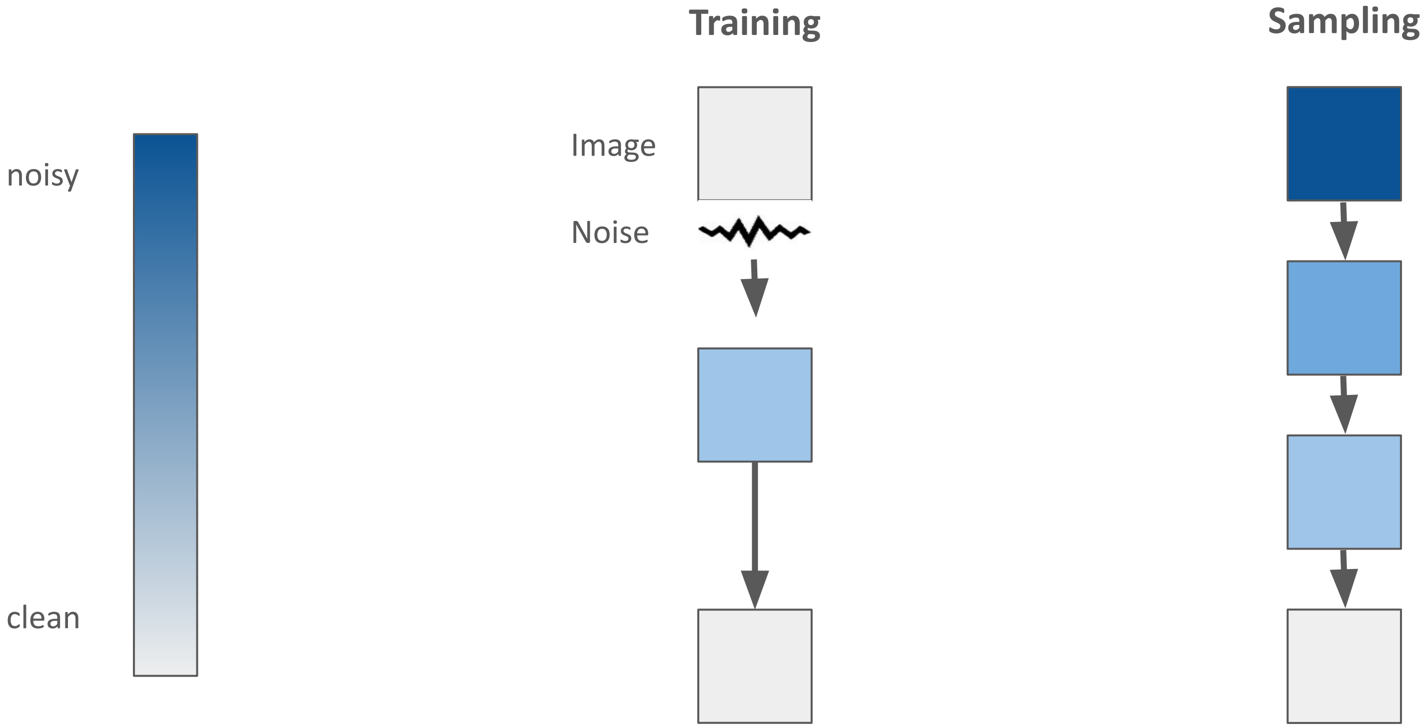
(-) Cannot guide the sampling to minimize a certain objective

(I think this is not true for diffusion models)

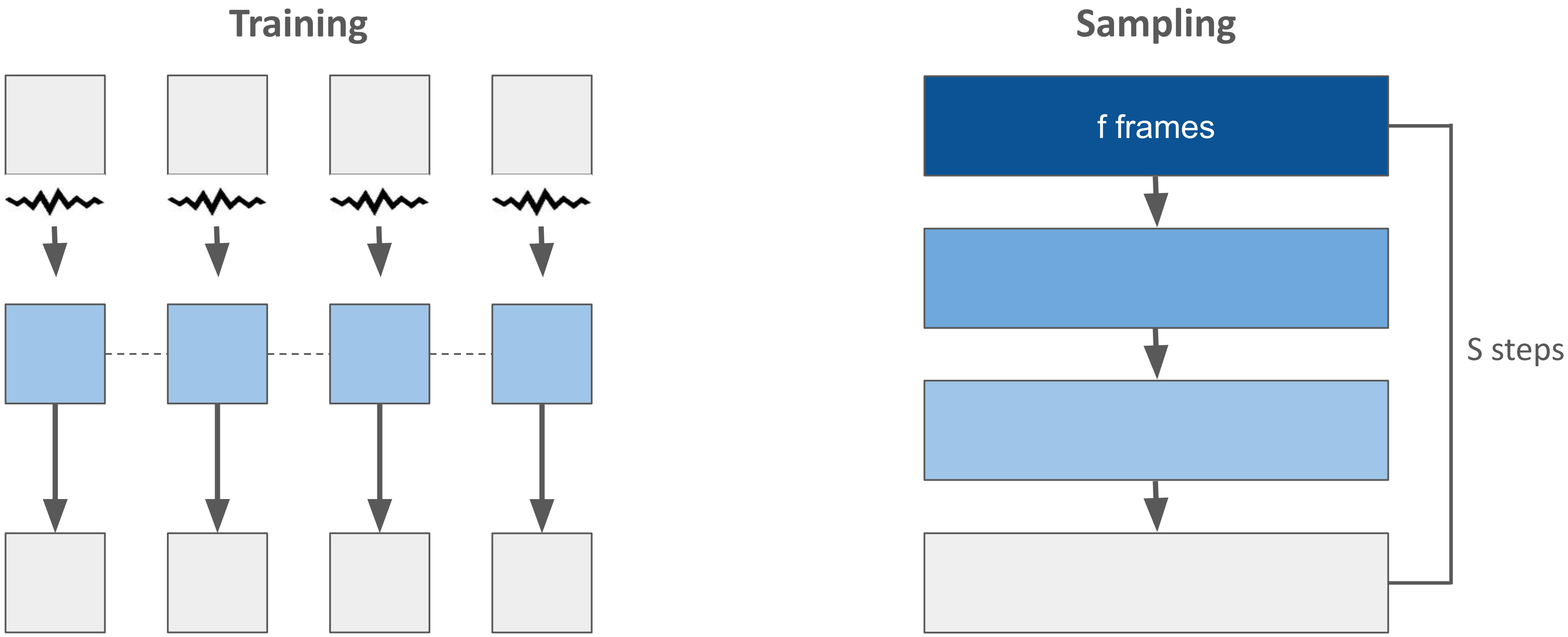


Full-Sequence Diffusion

Image Diffusion Models



Full-Sequence Diffusion



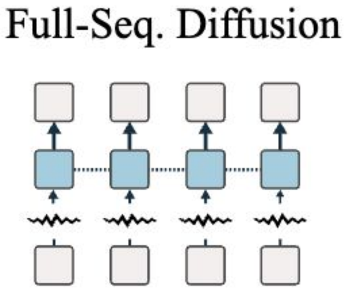
Full-Sequence Diffusion

Models denoise a fixed number of tokens with the **same noise level**.

(+) Guidance during iterative inference

(-) Non-causal modeling

(-) Limits in the number of generating frames

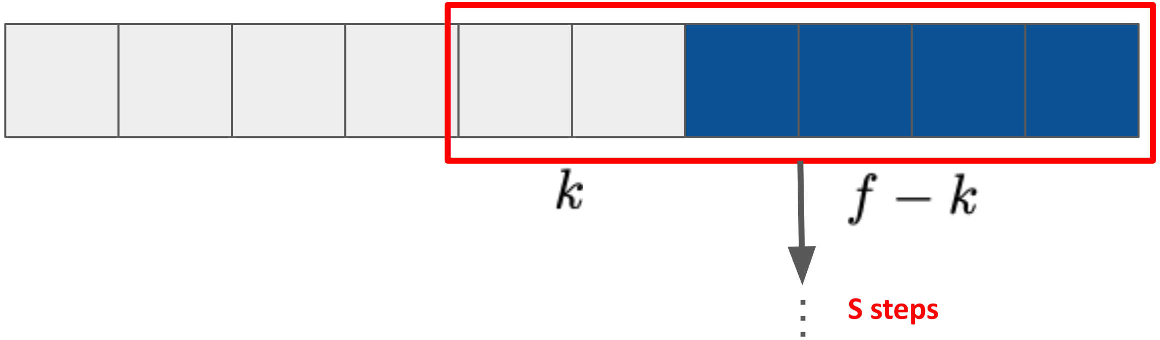


Long Video generation by Full-Sequence Diffusion Models

Chunked Autoregressive Generation

1. Generate f frames

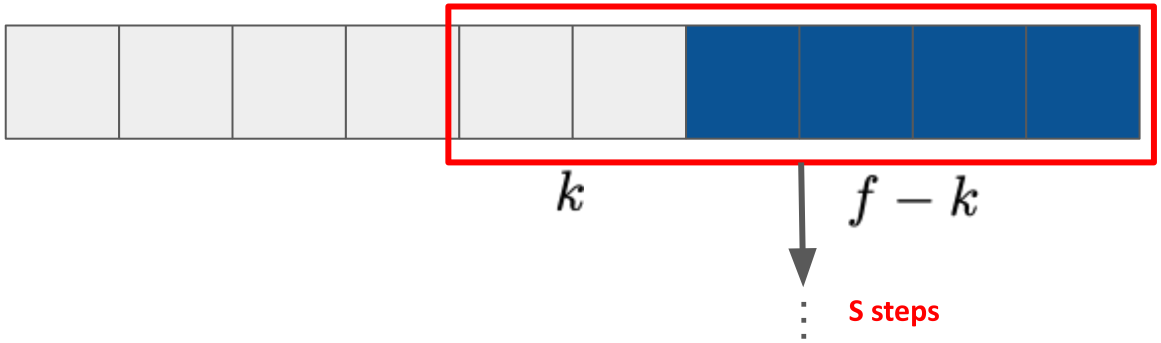
2. Conditioned on last k frames, denoise successive $f-k$ frames



Long Video generation by Full-Sequence Diffusion Models

Chunked Autoregressive Generation

2. Conditioned on last k frames, denoise successive $f - k$ frames



- Small k means **high latency** for each action.
- Large k (teacher forcing) means inefficient training and inference since token lengths are f, while models predict only **one token**.

Long Video generation by Full-Sequence Diffusion Models

Chunked Autoregressive Generation

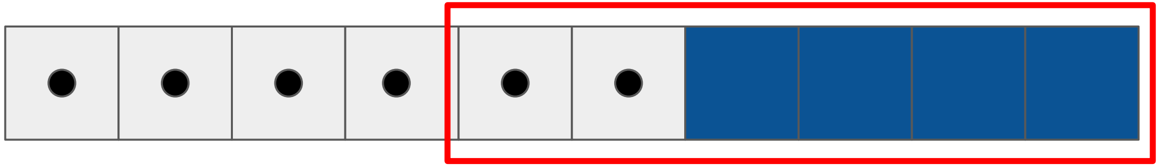
Problems

- Periodic discontinuity (small k)
- Quality degradation over time

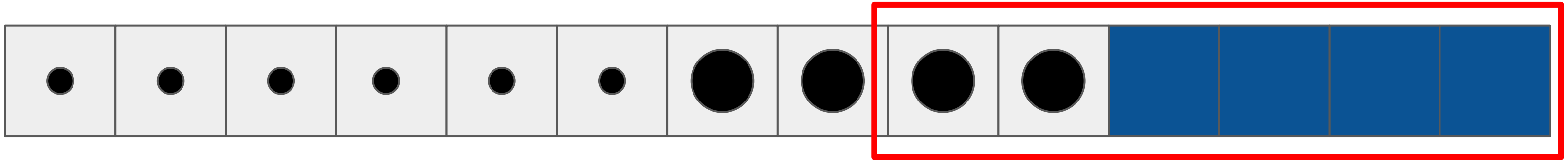
Long Video generation by Full-Sequence Diffusion Models

- **Error accumulation** causes quality degradations.

1. Generated samples have some error, but models do not know.



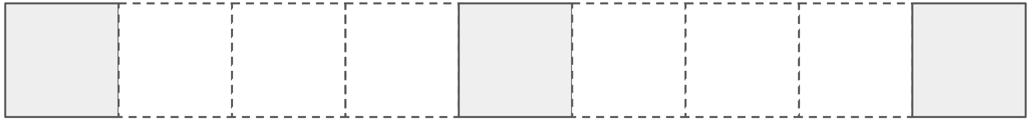
2. Successive frames have more error



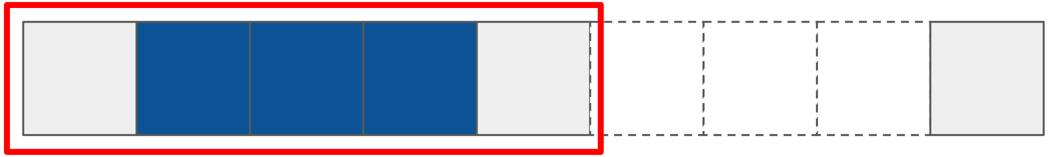
Generating Long Videos by Full-Sequence Diffusion Models

Hierarchical generation

1. Generate f key frames



2. Conditioned on key frames interpolate frames

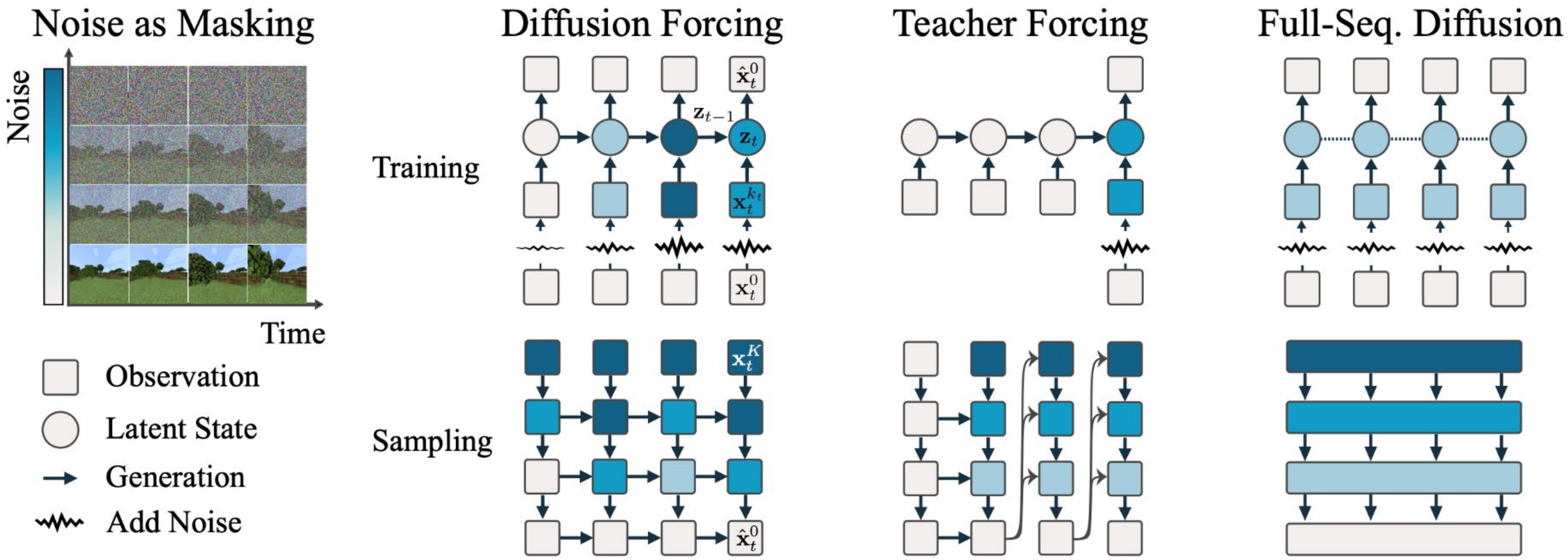


Full-Sequence Diffusion Models as World Simulator

	Chunked Autoregressive (k=1)	Chunked Autoregressive (k=f-1)	Hierarchical
Overview			
Problems as World Simulator	• High latency for each action • Forget context • Quality degradation over time	• Inefficiency in training and inference • Quality degradation over time	• Conceptually do not fit interactive simulator

Full-sequence diffusion models are not appropriate for world simulator!!

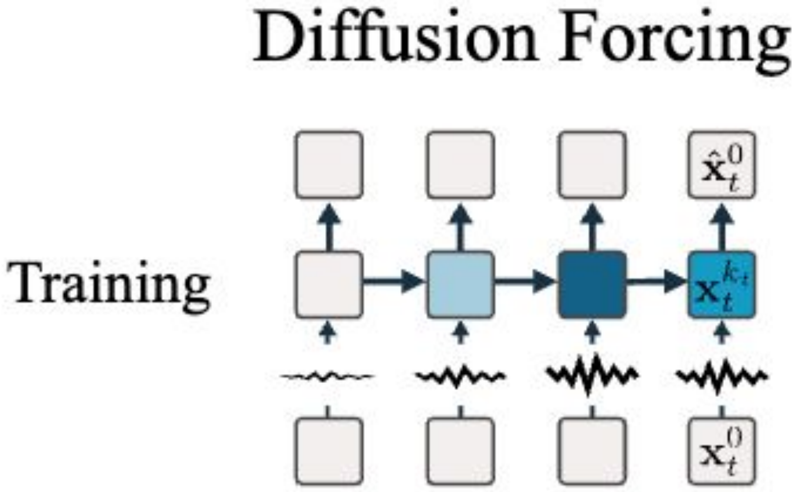
Diffusion Forcing = Teacher Forcing + Diffusion Models



Diffusion Forcing (Training)

Each token's noise level are independently sampled.

- Noising is considered as partial masking!



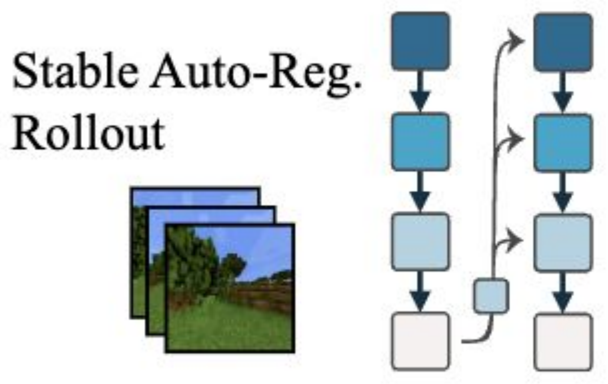
Diffusion Forcing (Training)

Computational overhead?

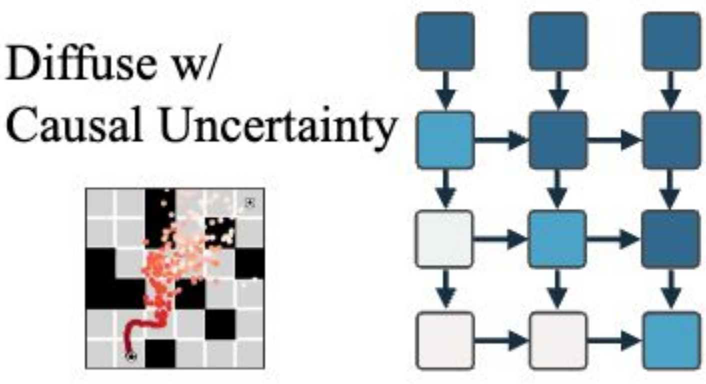
- Complexity added by independent noise level is in temporal dimension.
- Image pre-training, and then train on video data.

Diffusion Forcing (Inference)

Tokens are denoised following noise schedule. Noise schedule is dependent to inference purpose.



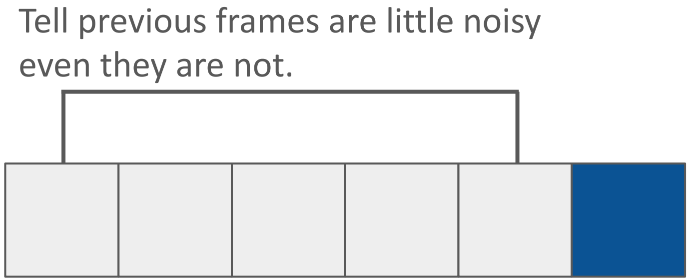
Similar to chunked autoregressive

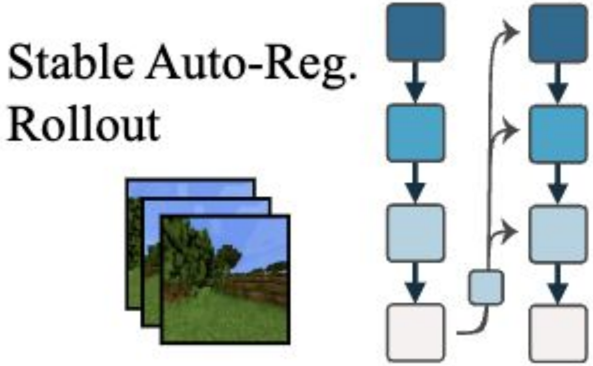


Similar to FIFO-Diffusion

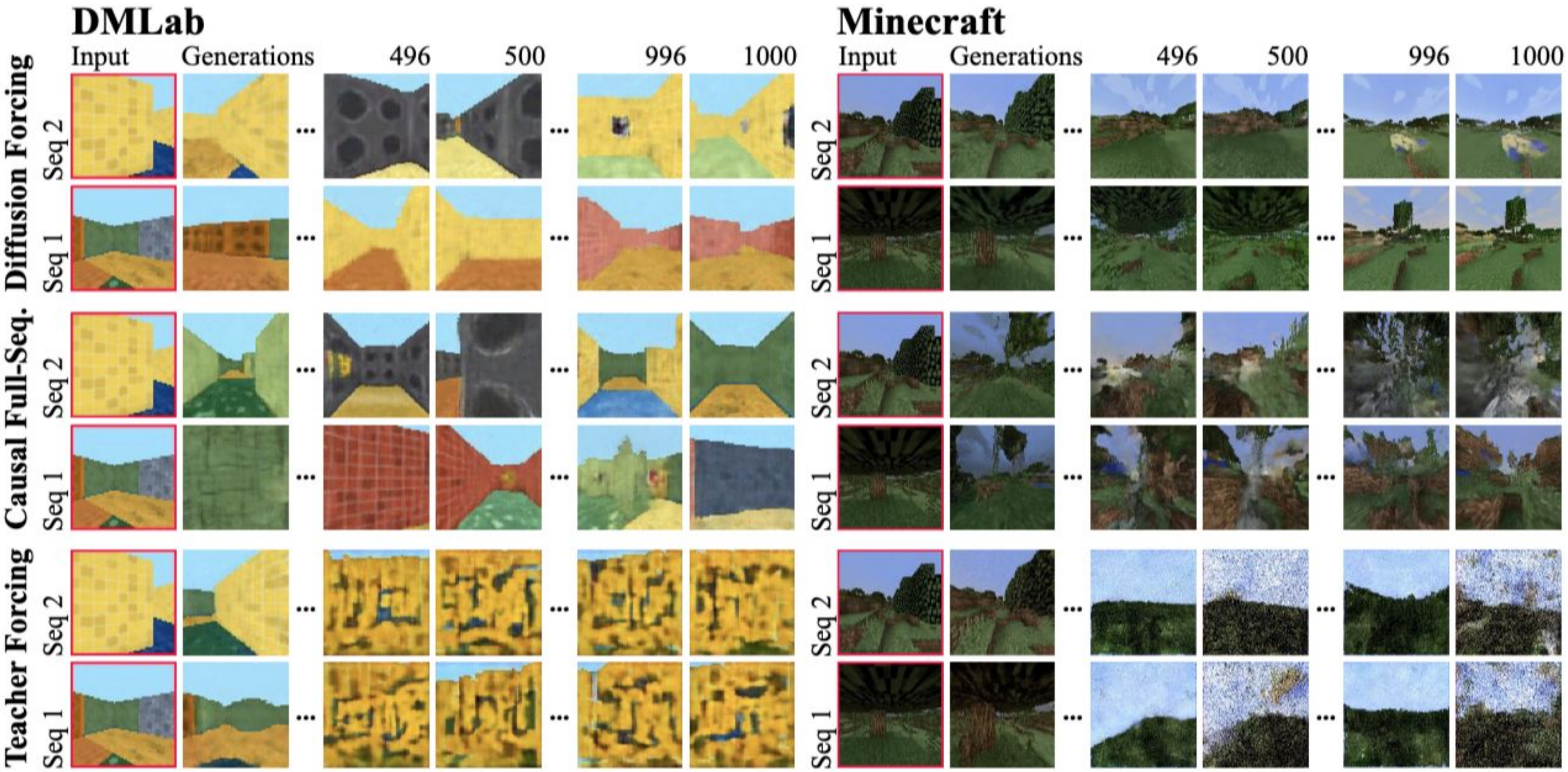
Diffusion Forcing for Long-Video Generation

- Chunked autoregressive methods suffer from quality degradation originated from error accumulation.
- To avoid it, Diffusion Forcing tells model that generated frames are little noisy even they are not.
  - Model will suspect generated frames may have error

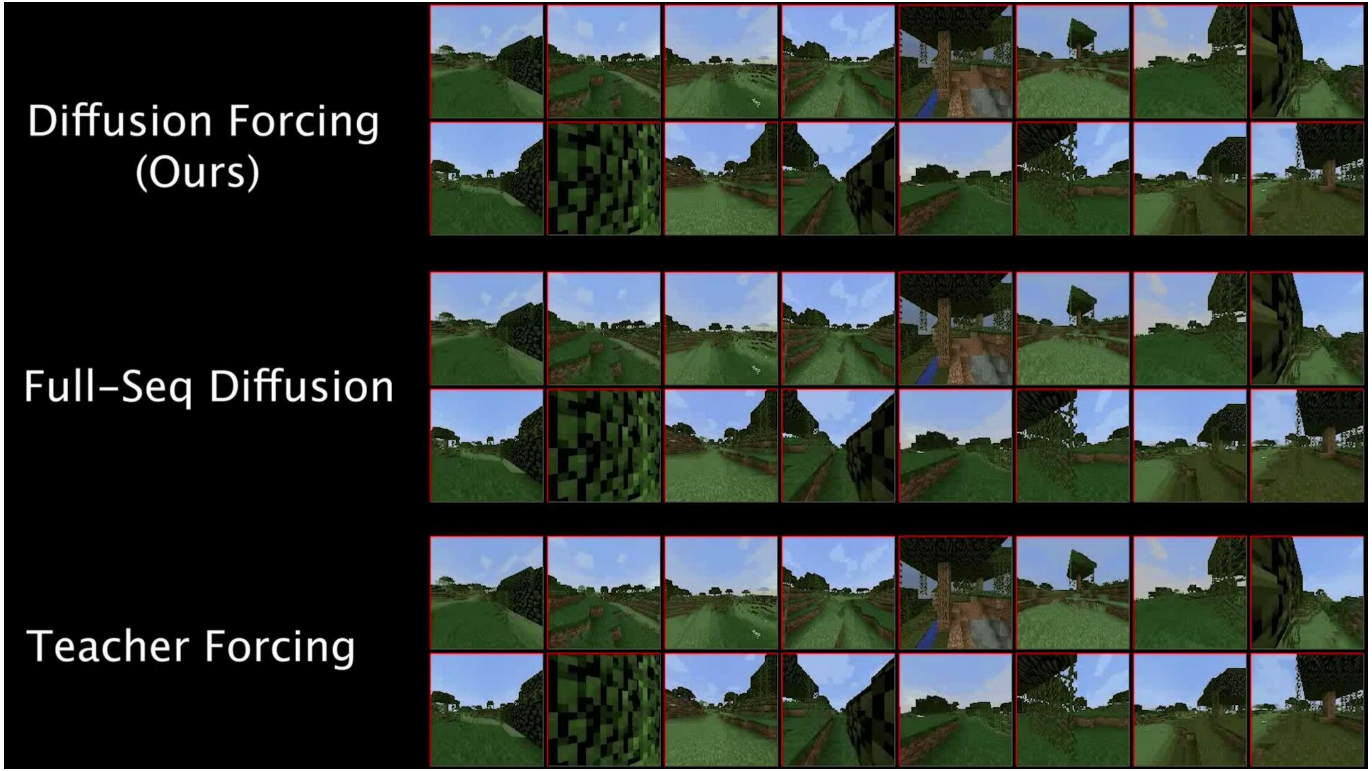




Diffusion Forcing for Long-Video Generation (RNN-based)



Diffusion Forcing for Long-Video Generation (RNN-based)



- Temporally consistent.
- Less quality degradation.

Contributions of Diffusion Forcing

- Independent noise levels
  - Stabilization of autoregressive rollout
  - For other purposes (causal uncertainty)
  - Cheaper training than training next-token prediction for full-sequence diffusion models.

- Stable autoregressive rollout (tell model that generated tokens are little noisy)
  - It is actually OOD
  - Seems little awkward

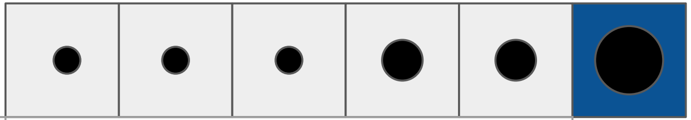
OASIS: A Universe in a Transformer

- OASIS uses transformer instead of RNN as Diffusion Forcing
    - Transformer is implemented with sliding window. Conceptually, it cannot remember whole history.

- OASIS has some modification on *stable autoregressive rollout*

OASIS: A Universe in a Transformer

Normal

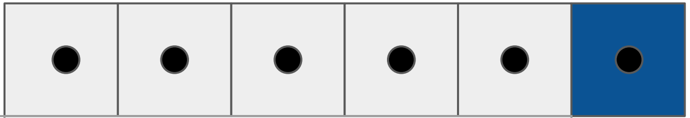


Actual noise level	Noise = 0
Tell models as	Noise = 0

- Consider noise level as model's trust in conditioning frames.
- Model believes that generated artifacts are GT, which leads to error accumulation.

OASIS: A Universe in a Transformer

Stable Rollout in Diffusion Forcing

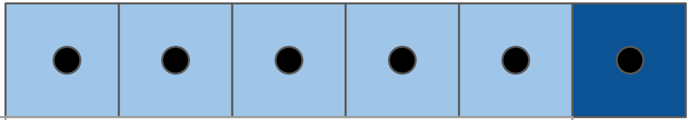


Actual noise level	Noise = 0
Tell models as	Noise = little noisy

- Model dost not fully trust pre-generated frames.
- Model will consider strange artifacts in pre-generated frames as noise, and prevent error accumulation.
- It is OOD.
- No magic number for “little noisy”

OASIS: A Universe in a Transformer

Stable rollout (Another option)

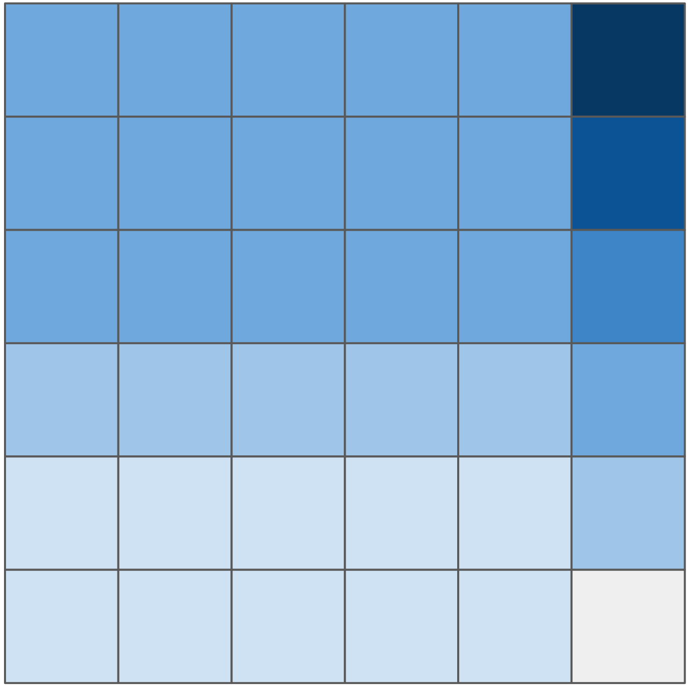


Actual noise level	Noise = little noisy
Tell models as	Noise = little noisy

- Model dost not fully trust pre-generated frames and model inputs are in distribution.
- Some details can be removed by adding noise.

OASIS: A Universe in a Transformer

Dynamic Noise Augmentation



- For initial denoising steps, add moderate noise to conditioning tokens.
  - At initial steps, models generate low-frequency features. Therefore, it is okay to lose some details by adding noise.
- For last denoising steps, noise levels of conditioning tokens gradually decreases.
  - Artifacts cannot grow.

OASIS: A Universe in a Transformer

Still there are many limitations of the models. For example, models forget history.

Limited memory over long horizons

OASIS: A Universe in a Transformer

They somewhat solved memory problem for short time horizon.

OASIS: A Universe in a Transformer

Their next step? Long-Term Memory

- Adaptive Memory
  - Pick frames to refer to dynamically.
- Mixed-SSMs
- Spatially-aware memory
  - 3D representation makes sense