SEQUENTIAL LATENT VARIABLE MODELS & FILTERING

JOSEPH MARINO
CALTECH

OUTLINE

- sequence models
- amortized variational filtering
- sequential autoregressive flows

SEQUENCE MODELS

GENERATIVE MODEL

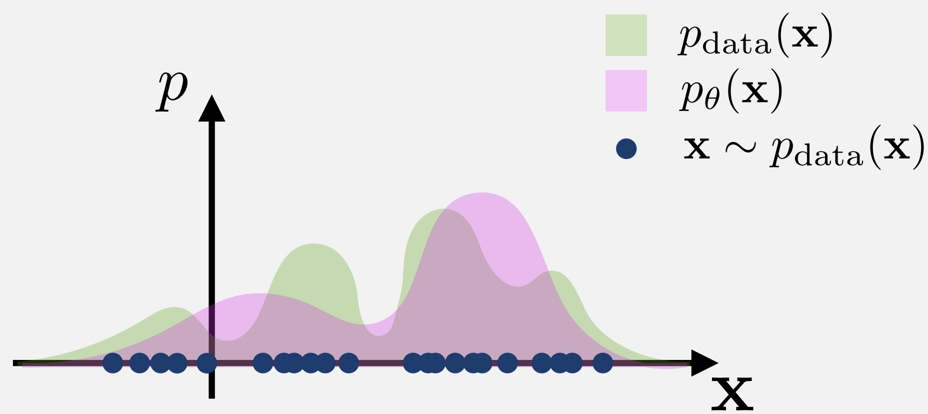
a model of the density of observed data

MAXIMUM LIKELIHOOD

data: $p_{\text{data}}(\mathbf{x})$

model: $p_\theta(\mathbf{x})$

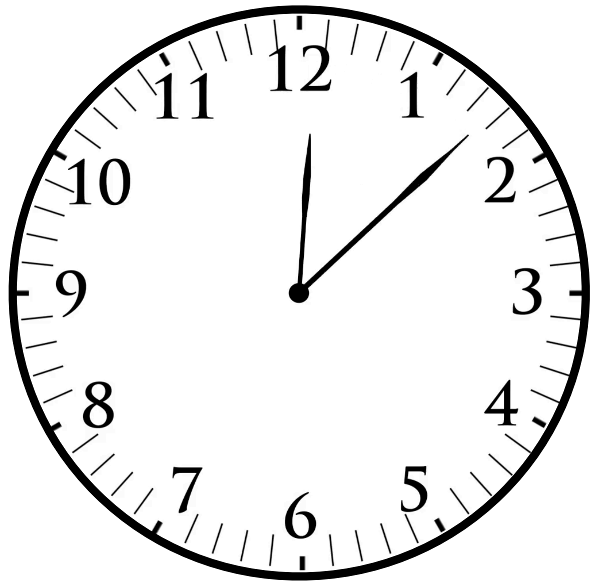
parameters: θ



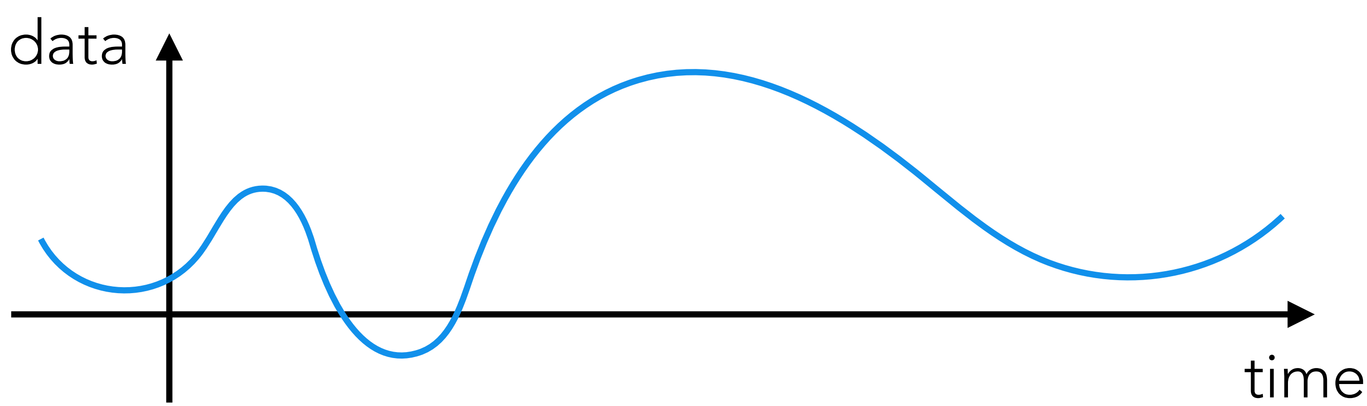
maximum likelihood estimation

find the model that assigns the *maximum likelihood* to the data

$$\theta^* = \arg\max_\theta \; \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left[\log p_\theta(\mathbf{x})\right] \approx \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)})$$

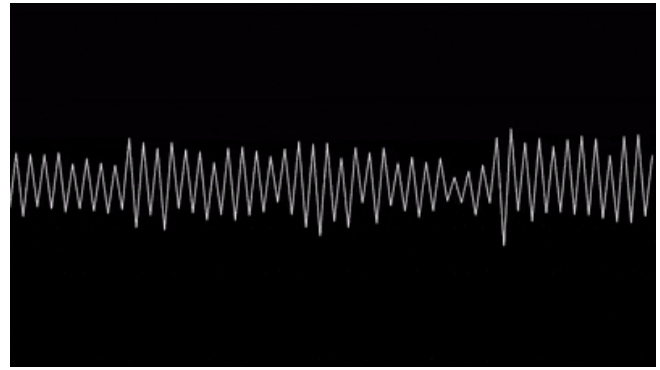


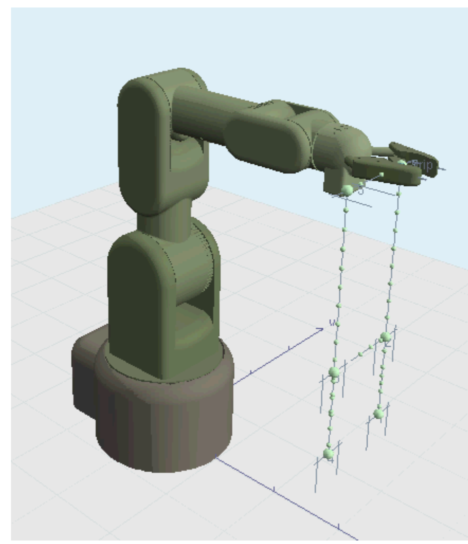
observed data are often sequential



vision

audio



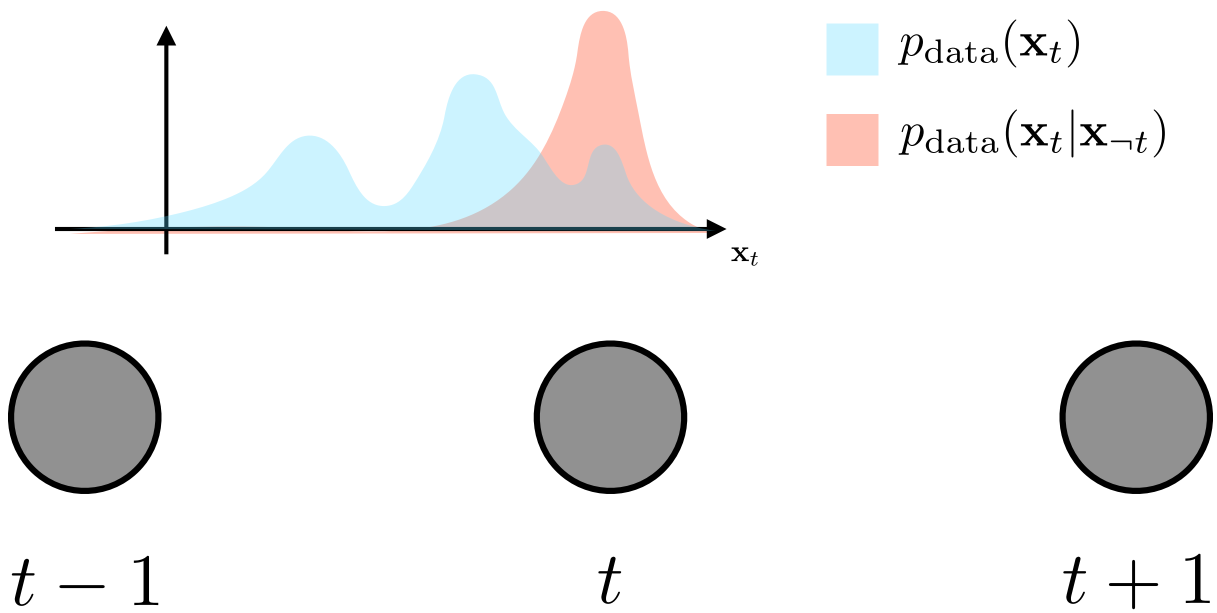


joint angles

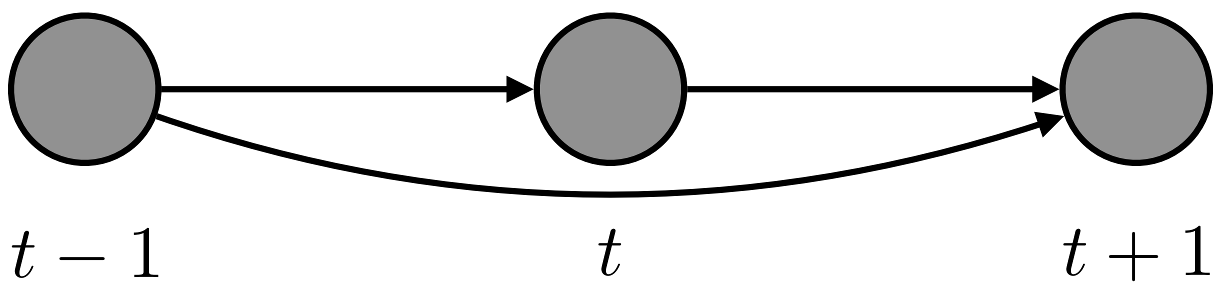
dynamics: dependence in time

multi-information: $\mathcal{I}(\mathbf{x}_{1:T}) = \sum_t \mathcal{H}(\mathbf{x}_t) - \mathcal{H}(\mathbf{x}_{1:T}) \geq 0$

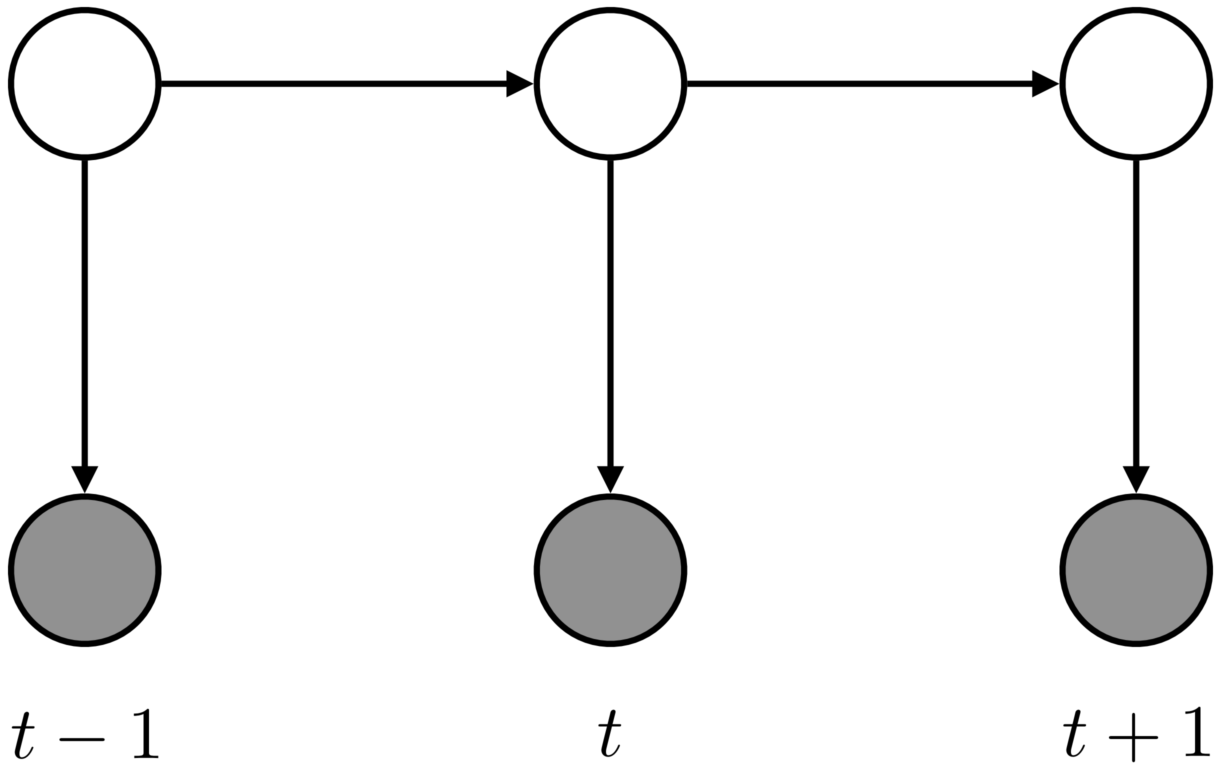
observing $\mathbf{x}_{\neg t}$ *reduces uncertainty* in $\mathbf{x}_t$



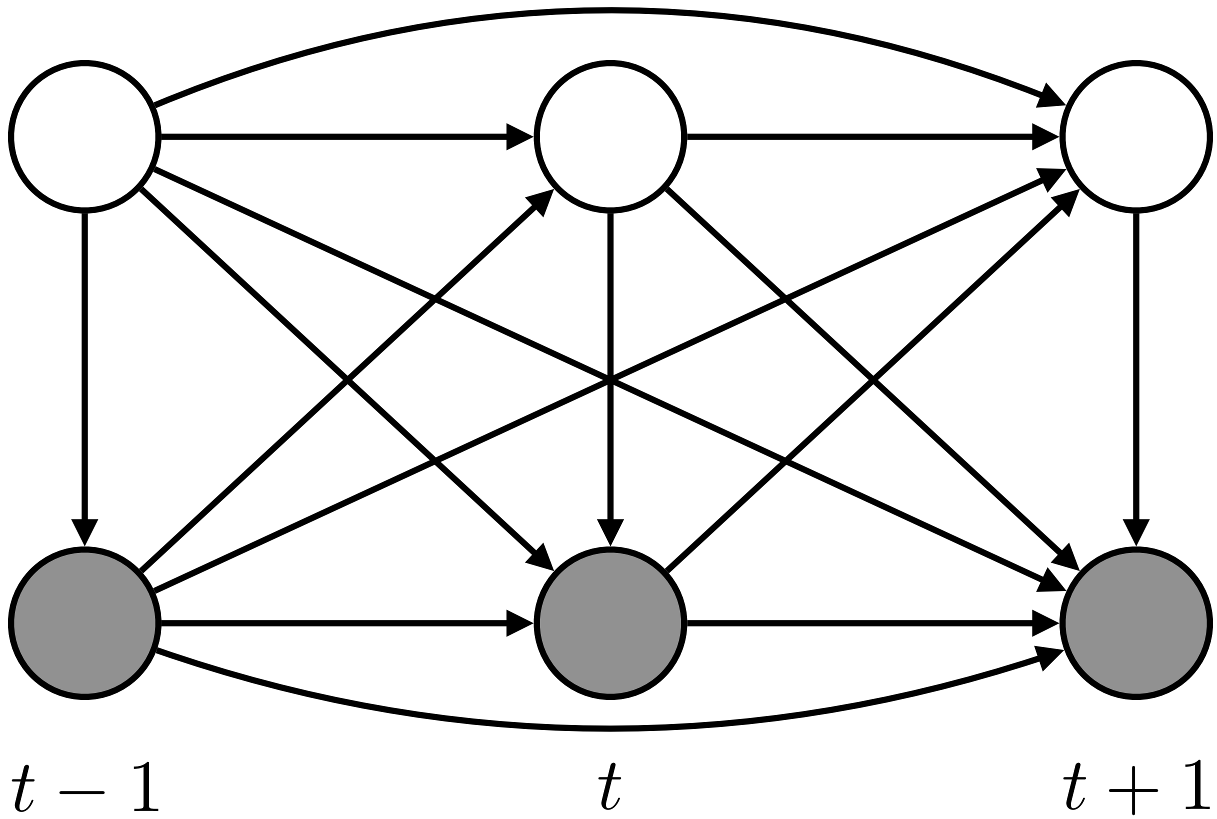
model temporal dependencies



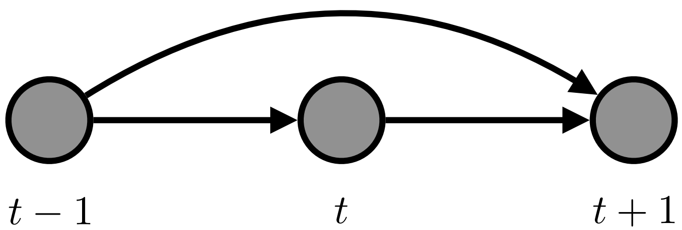
model temporal dependencies



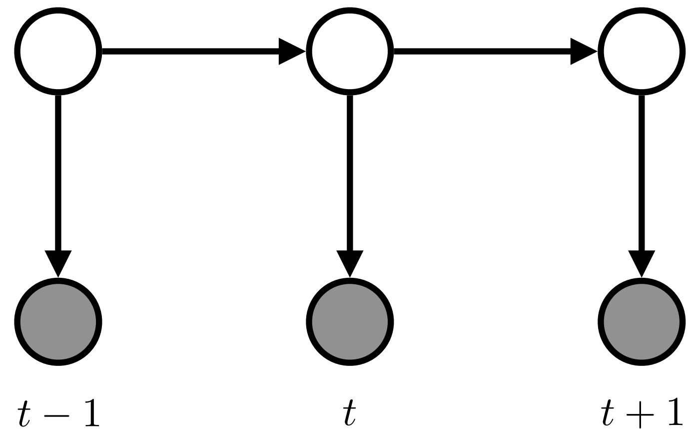
model temporal dependencies



MODELING DYNAMICS



fully-observed



latent

$$p_\theta(\mathbf{x}_t|\mathbf{x}_{<t}) = \int p_\theta(\mathbf{x}_t|\mathbf{z}_t) p_\theta(\mathbf{z}_t|\mathbf{x}_{<t}) d\mathbf{z}_t$$

mixture component — mixture probability

may be more flexible than a fixed-form $p_\theta(\mathbf{x}_t|\mathbf{x}_{<t})$

SEQUENTIAL LATENT VARIABLE MODELS

general form:

$$p_\theta(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) = \prod_{t=1}^{T} \underbrace{p_\theta(\mathbf{x}_t | \mathbf{x}_{<t}, \mathbf{z}_{\leq t})}_{\text{likelihood}} \underbrace{p_\theta(\mathbf{z}_t | \mathbf{x}_{<t}, \mathbf{z}_{<t})}_{\text{prior}}$$

where $\mathbf{x}_{\leq T}$ is a sequence of T observed variables

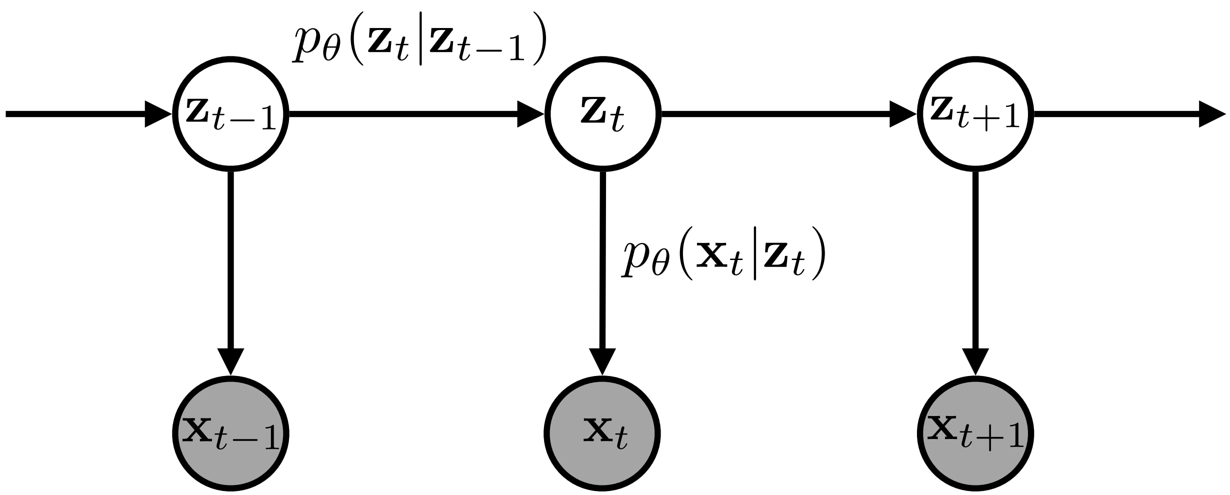
$\mathbf{z}_{\leq T}$ is a sequence of T latent variables

SEQUENTIAL LATENT VARIABLE MODELS

general form:

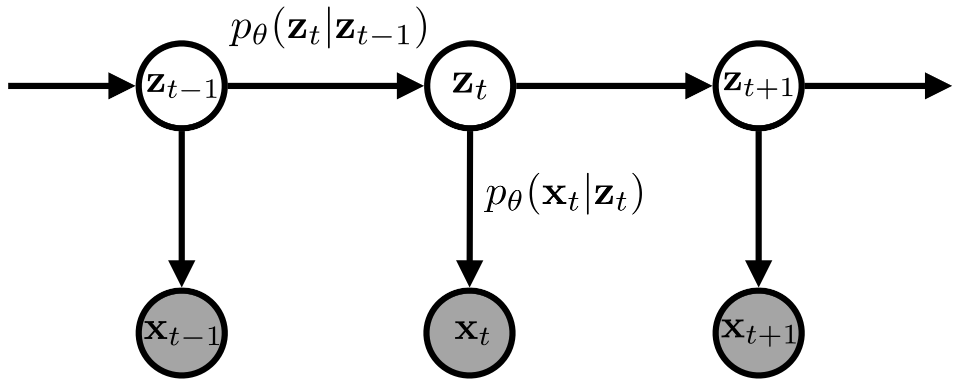
$$p_\theta(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) = \prod_{t=1}^{T} \underbrace{p_\theta(\mathbf{x}_t | \mathbf{x}_{<t}, \mathbf{z}_{\leq t})}_{\text{likelihood}} \underbrace{p_\theta(\mathbf{z}_t | \mathbf{x}_{<t}, \mathbf{z}_{<t})}_{\text{prior}}$$

simplified case (hidden Markov model):



SEQUENTIAL LATENT VARIABLE MODELS

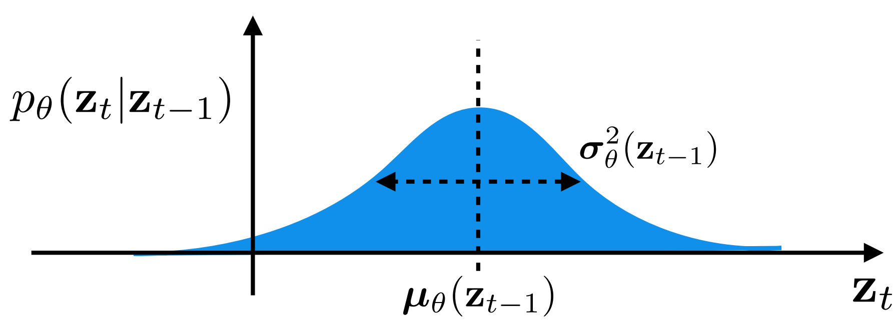
Markov model:



Parameterization:

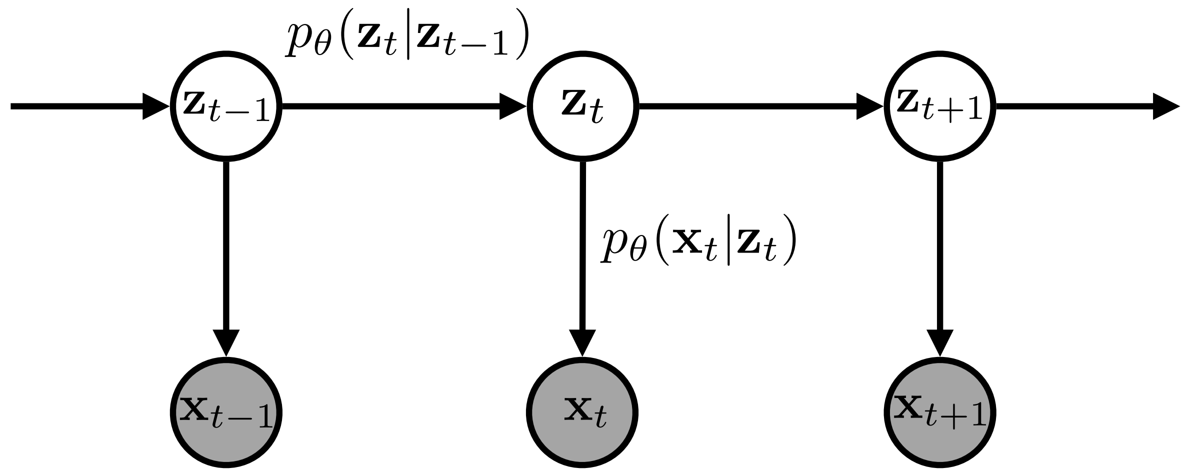
$p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1})$ is typically an analytical distribution

for example, $p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_\theta(\mathbf{z}_{t-1}), \text{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z}_{t-1})))$



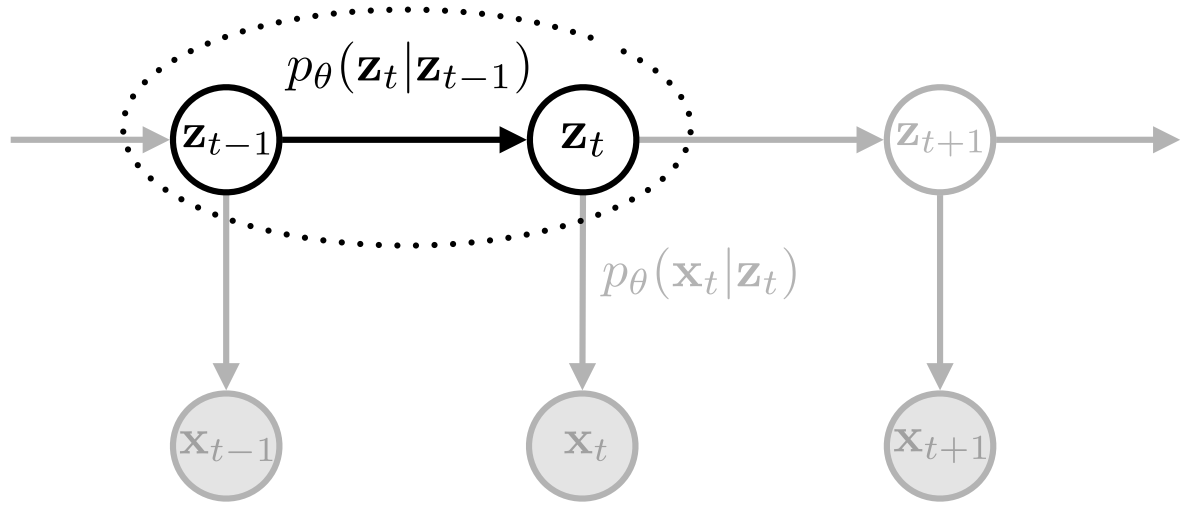
SEQUENTIAL LATENT VARIABLE MODELS

the parameters of these analytical distributions are functions, often *deep networks*



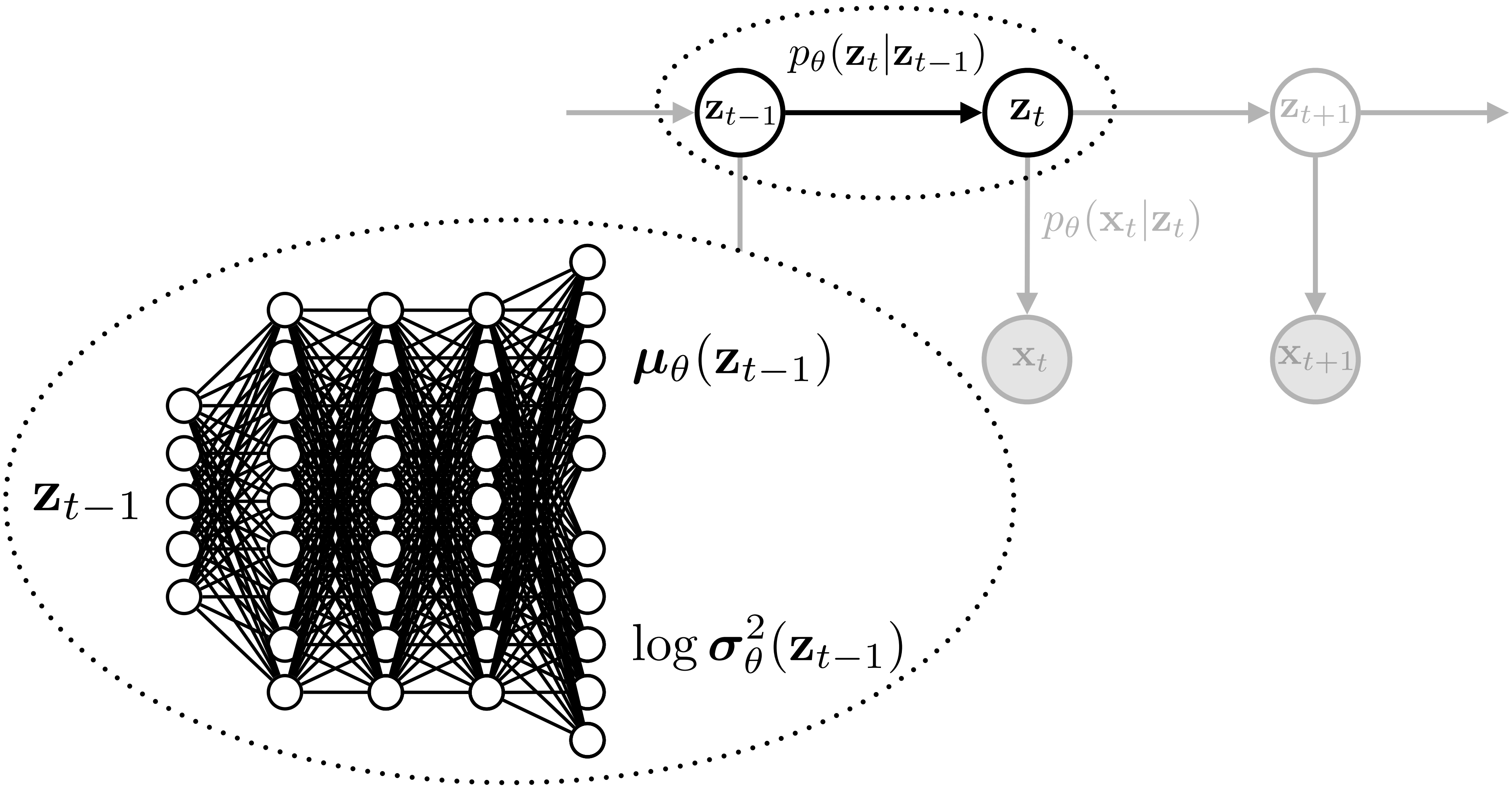
SEQUENTIAL LATENT VARIABLE MODELS

the parameters of these analytical distributions are functions, often *deep networks*



SEQUENTIAL LATENT VARIABLE MODELS

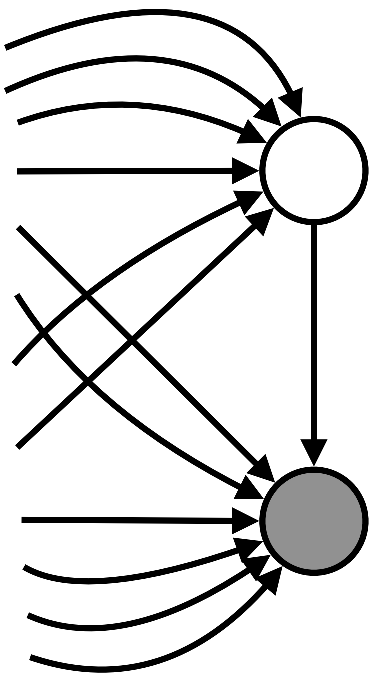
the parameters of these analytical distributions are functions, often *deep networks*



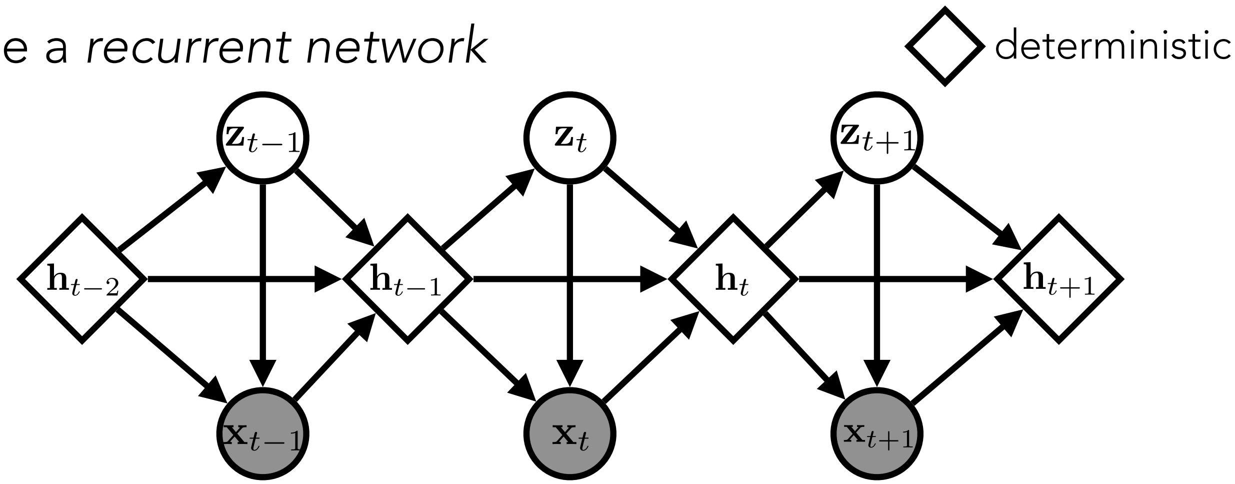
LONG-TERM DEPENDENCIES

general model form $p_\theta(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t}) p_\theta(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t})$

how do we model long-term dependencies?



use a *recurrent network*

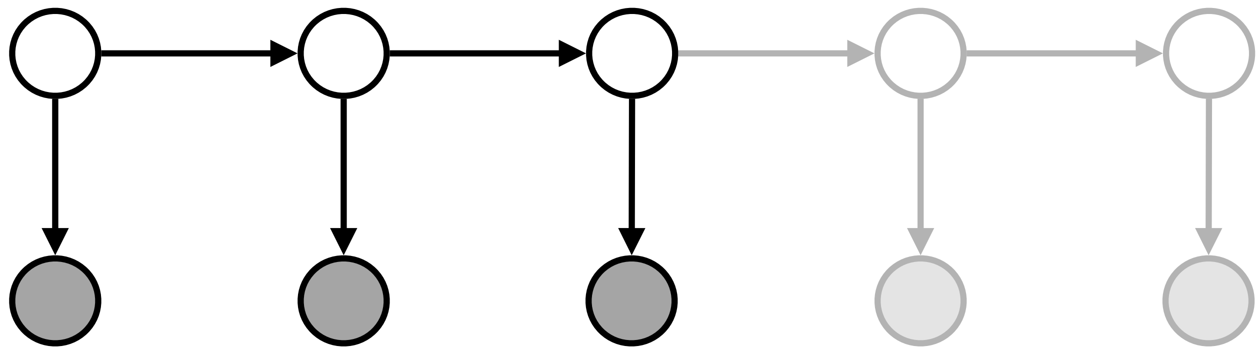


$$p_\theta(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{h}_{t-1}, \mathbf{z}_t) p_\theta(\mathbf{z}_t|\mathbf{h}_{t-1})$$

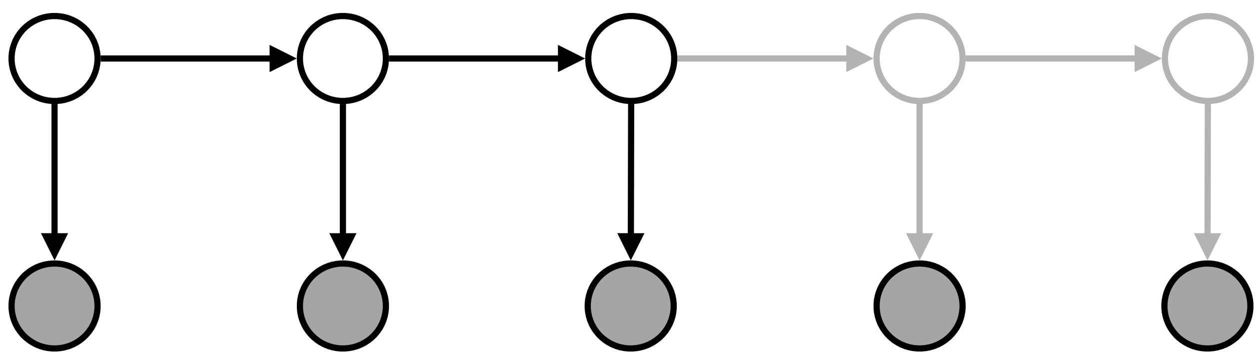
INFERENCE

given a sequence of observations, $\mathbf{x}_{\leq T}$, infer $p_\theta(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})$

filtering inference



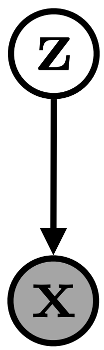
smoothing inference



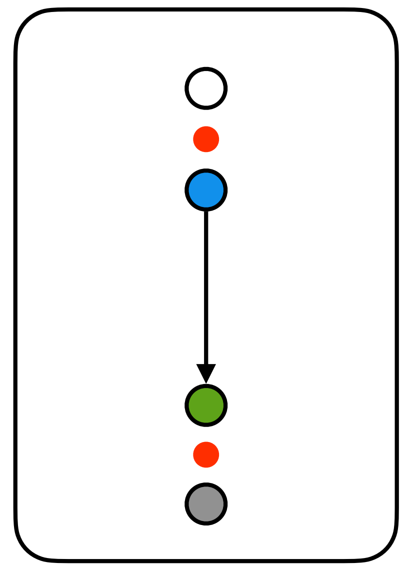
ASIDE: VARIATIONAL INFERENCE

VARIATIONAL INFERENCE

graphical model



computation graph



$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})$$

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

intractable

VARIATIONAL INFERENCE

approximate posterior $q(\mathbf{z}|\mathbf{x})$

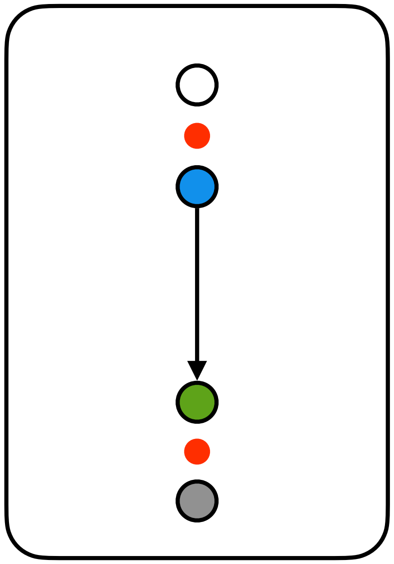
variational lower bound

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{"reconstruction"}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{"regularization"}} \right]$$

computation graph



VARIATIONAL INFERENCE

approximate posterior $q(\mathbf{z}|\mathbf{x})$

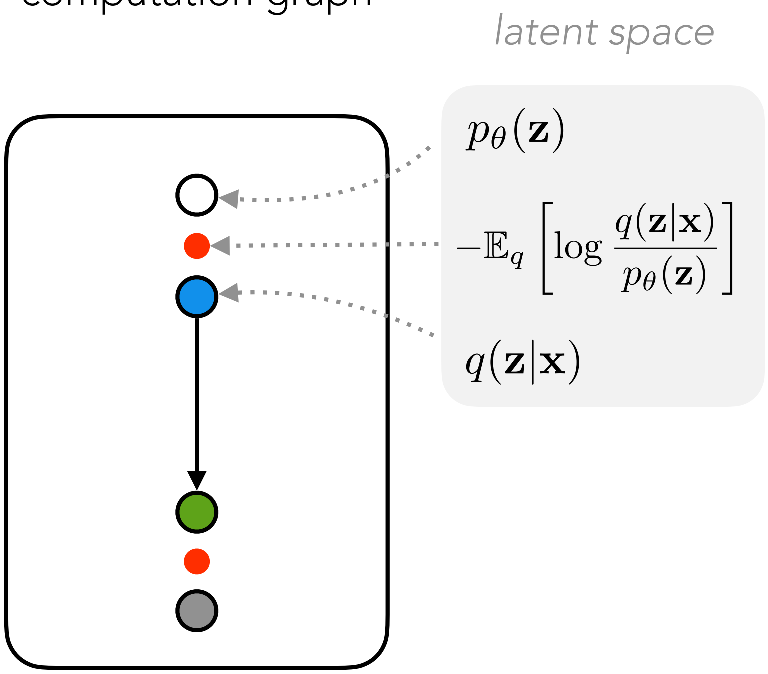
variational lower bound

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{"reconstruction"}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{"regularization"}} \right]$$

computation graph



VARIATIONAL INFERENCE

approximate posterior $q(\mathbf{z}|\mathbf{x})$

---

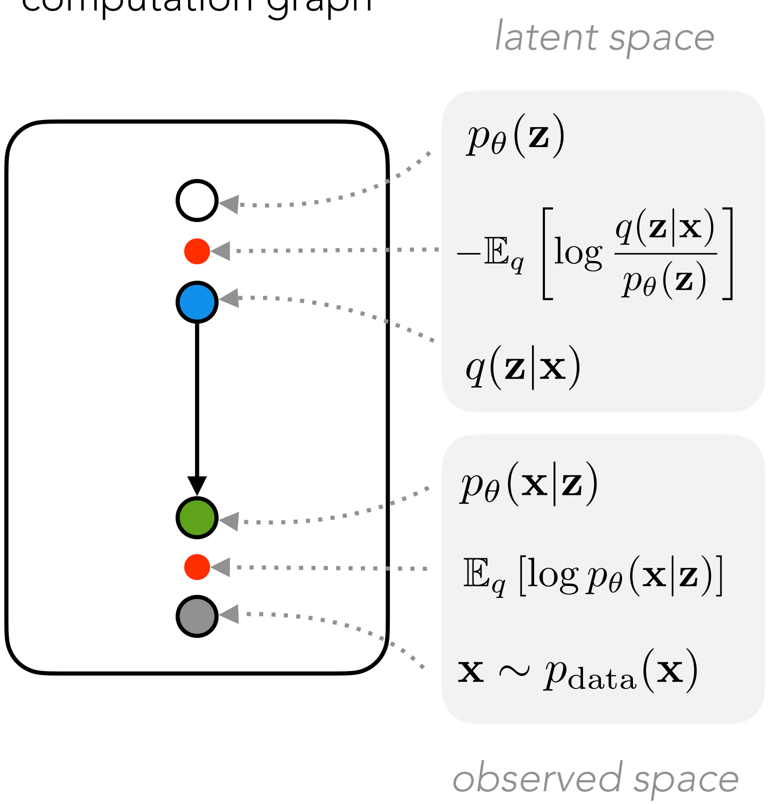
variational lower bound

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{"reconstruction"}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{"regularization"}} \right]$$

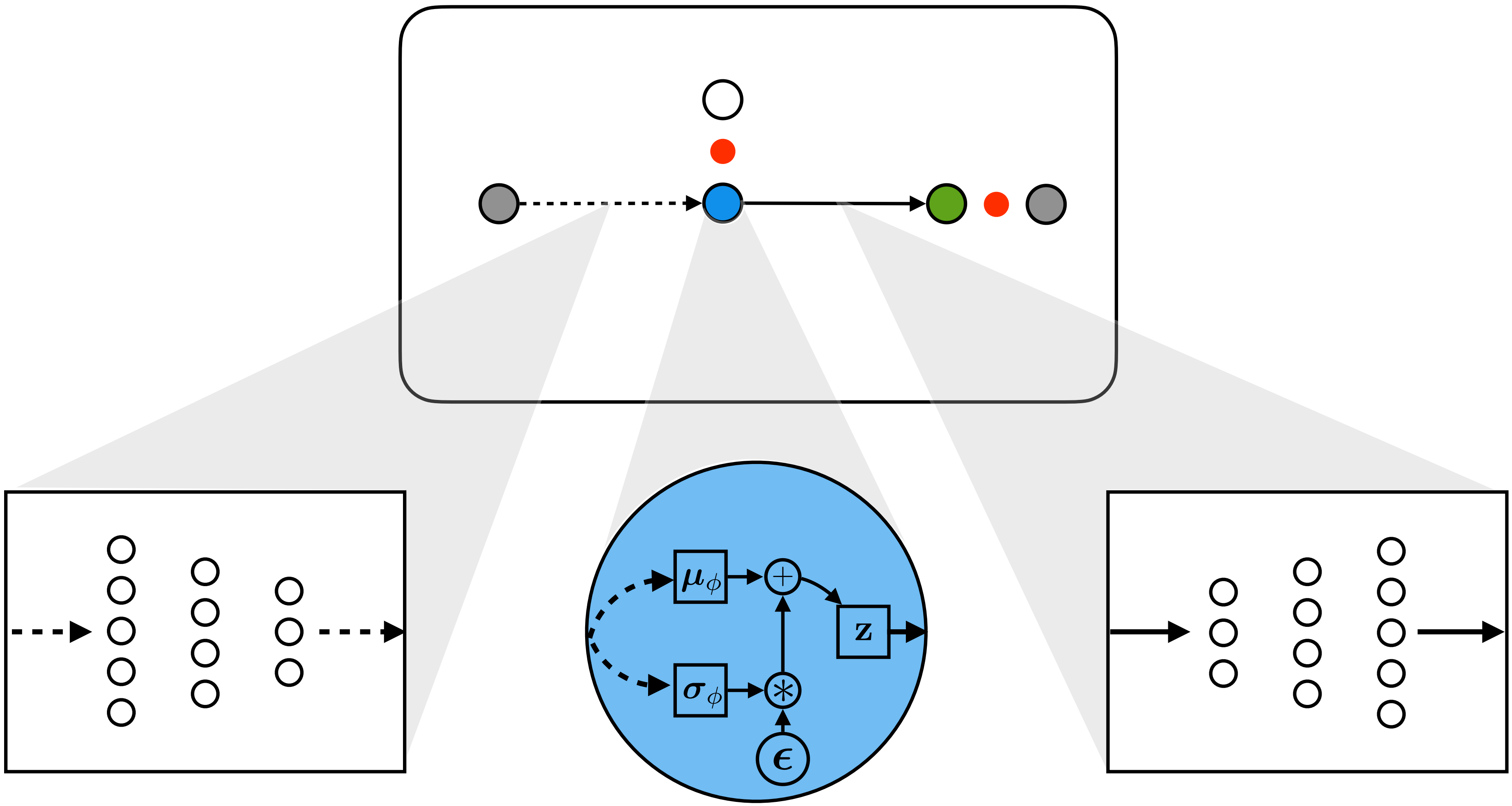
computation graph



AMORTIZED INFERENCE

Variational Autoencoder (VAE):

deep latent variable model + variational inference + direct encoder + *reparameterized* Gaussian



Kingma & Welling, 2014
Rezende et al., 2014

VARIATIONAL INFERENCE IN SEQUENTIAL MODELS

introduce an approximate posterior $q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})$

ELBO: $$\mathcal{L}(\mathbf{x}_{\leq T}, q) = \mathbb{E}_q \left[\log \frac{p_\theta(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T})}{q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})} \right]$$

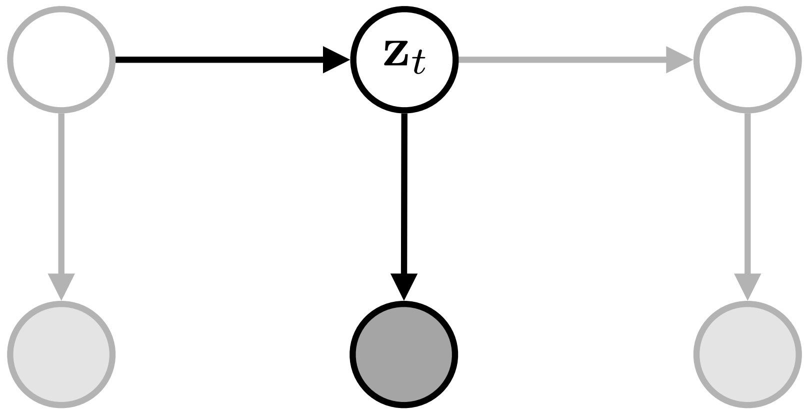
choices about the form of $q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})$ determine how we evaluate $\mathcal{L}$

⟶ often $q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T})$ is *structured*

STRUCTURED VARIATIONAL INFERENCE

the model contains temporal dependencies

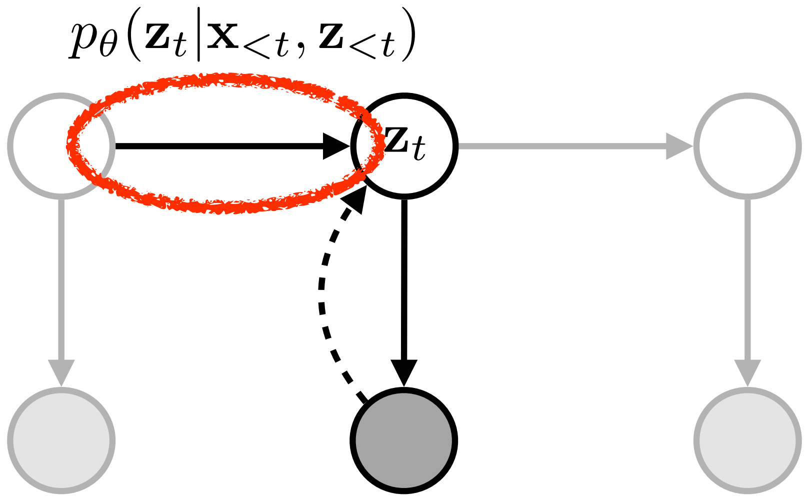
$$p_\theta(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t})$$



the approximate posterior should account for these dependencies

STRUCTURED VARIATIONAL INFERENCE

the model contains temporal dependencies



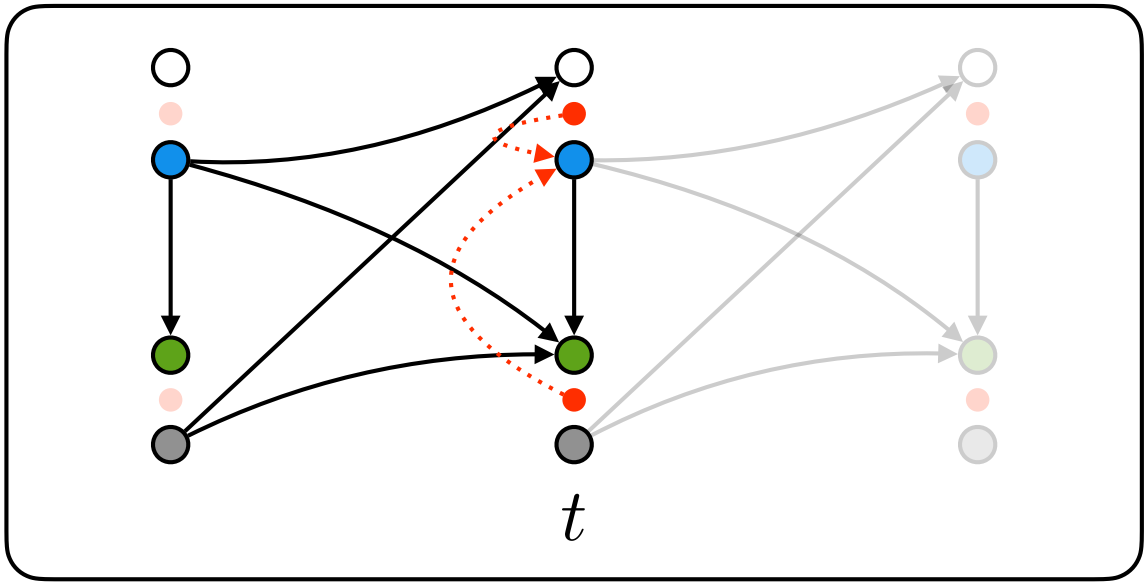
the approximate posterior should account for these dependencies

→ if we use $q(\mathbf{z}_t | \mathbf{x}_t)$, we cannot account for $\mathbf{x}_{<t}$ and $\mathbf{z}_{<t}$

FILTERING INFERENCE

filtering approximate posterior

$$q(\mathbf{z}_{\leq T}|\mathbf{x}_{\leq T}) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$$



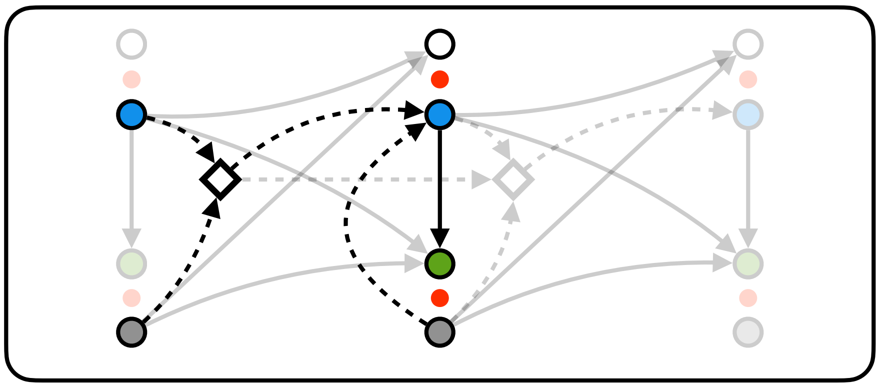
condition on observations at past and present time steps

AMORTIZED VARIATIONAL INFERENCE

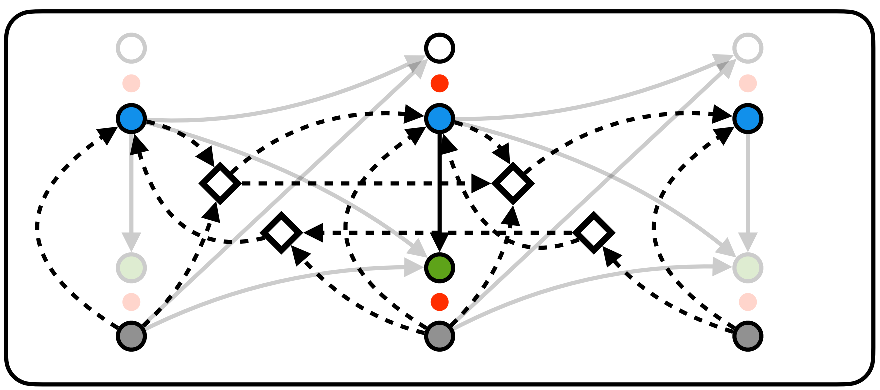
how do we amortize inference in sequential models?

typical approach:

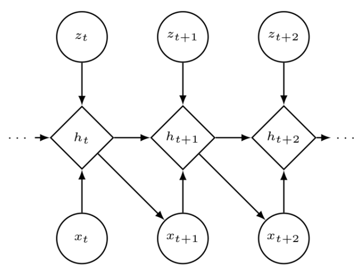
filtering: use a recurrent network



smoothing: use a bi-directional recurrent network

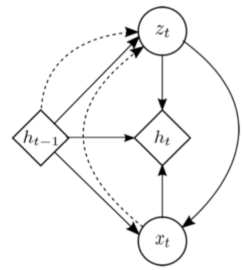


MODELS



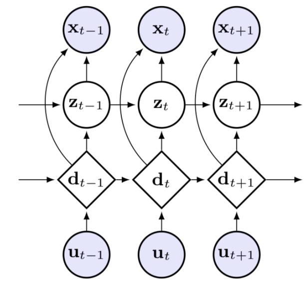
STORN

Bayer & Osendorfer, 2014



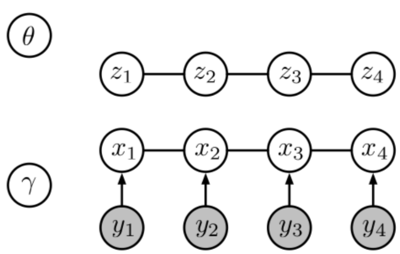
VRNN

Chung *et al.*, 2015



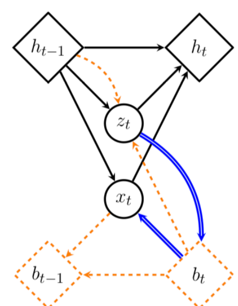
SRNN

Fraccaro *et al.*, 2016



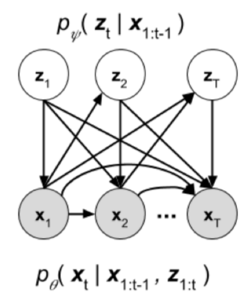
Structured VAE

Johnson *et al.*, 2016



Z-Forcing

Goyal *et al.*, 2017

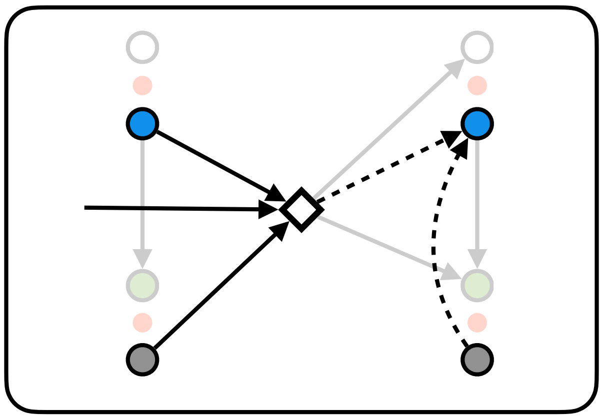


SVG

Denton & Fergus, 2018

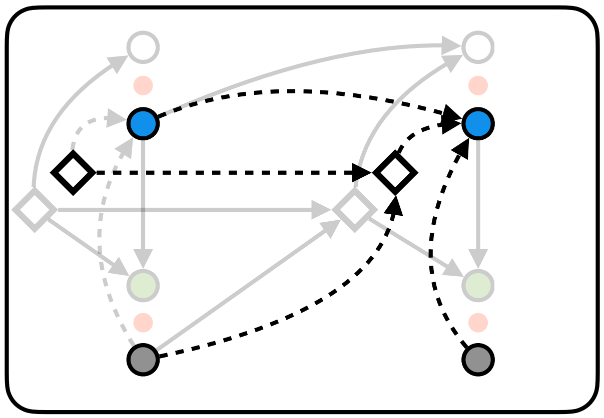
• • •

FILTERING INFERENCE MODELS



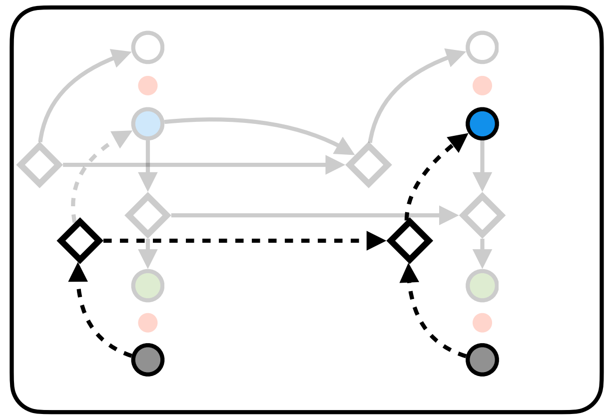
VRNN

Chung et al., 2015



SRNN

Fraccaro et al., 2016



SVG

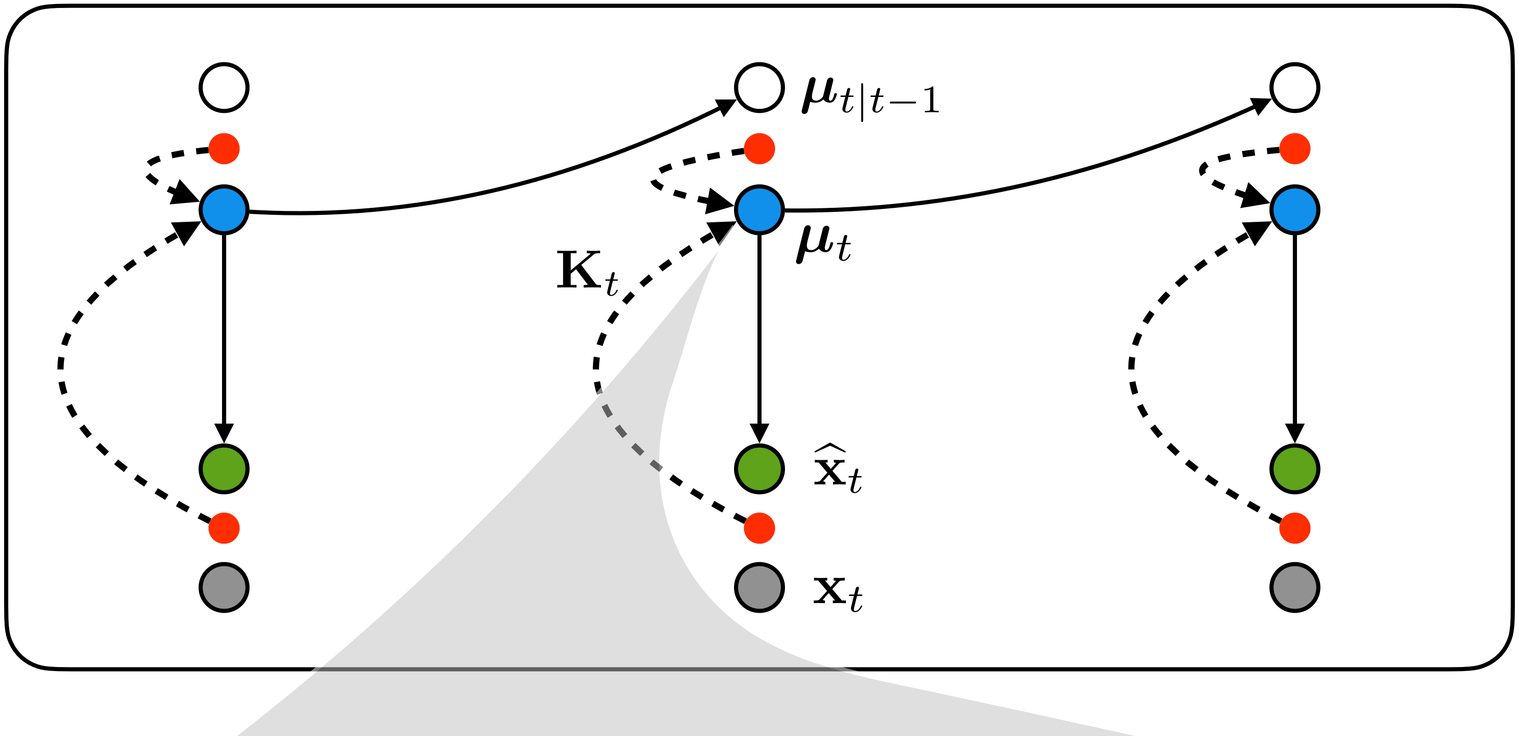
Denton & Fergus, 2018

custom-designed inference models

AMORTIZED VARIATIONAL FILTERING

KALMAN FILTERING

Kalman filtering: exact Bayesian inference in linear-Gaussian model



$$\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{x}_t - \widehat{\mathbf{x}}_t)$$

estimate $\leftarrow$ prediction + gain $\cdot$ prediction error

ITERATIVE AMORTIZED INFERENCE

let $\boldsymbol{\lambda}$ be the distribution parameters of $q(\mathbf{z}|\mathbf{x})$, for example, $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$

inference optimization: $q(\mathbf{z}|\mathbf{x}) \leftarrow \arg\max_q \mathcal{L}(\mathbf{x}; q)$

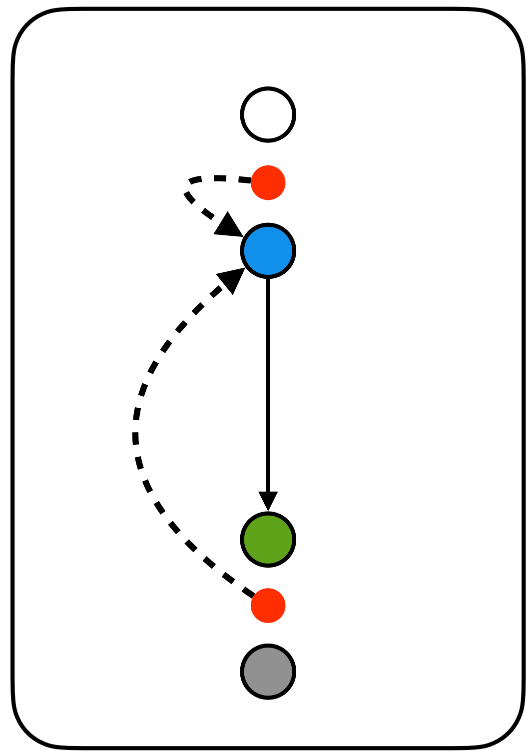
ITERATIVE AMORTIZED INFERENCE

learn an iterative mapping

$$\boldsymbol{\lambda} \leftarrow f_\phi(\boldsymbol{\lambda}, \nabla_{\boldsymbol{\lambda}} \mathcal{L})$$

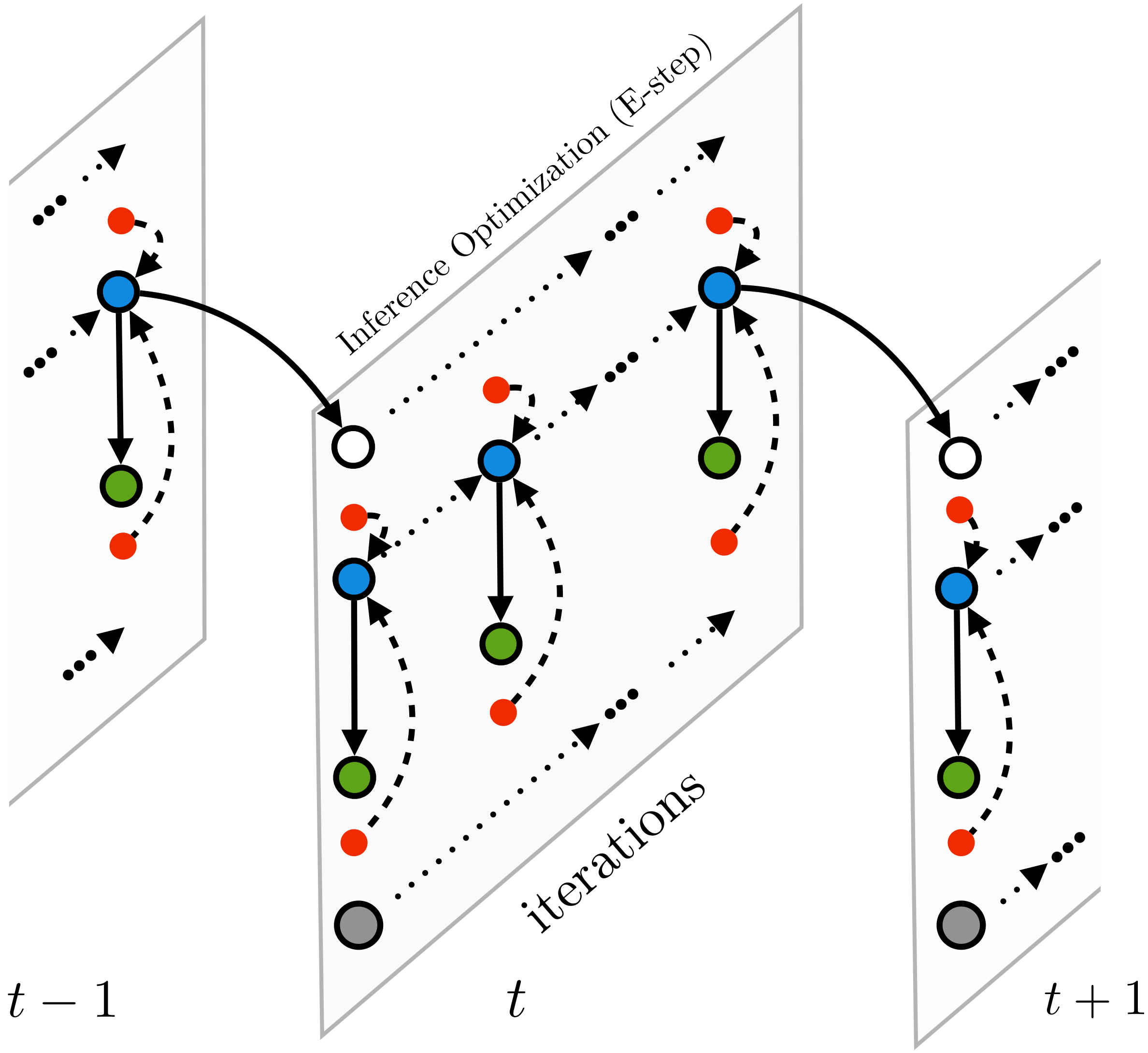
contains prediction errors

$\mathbf{x} - \widehat{\mathbf{x}}$



Marino *et al.*, 2018a

AMORTIZED VARIATIONAL FILTERING

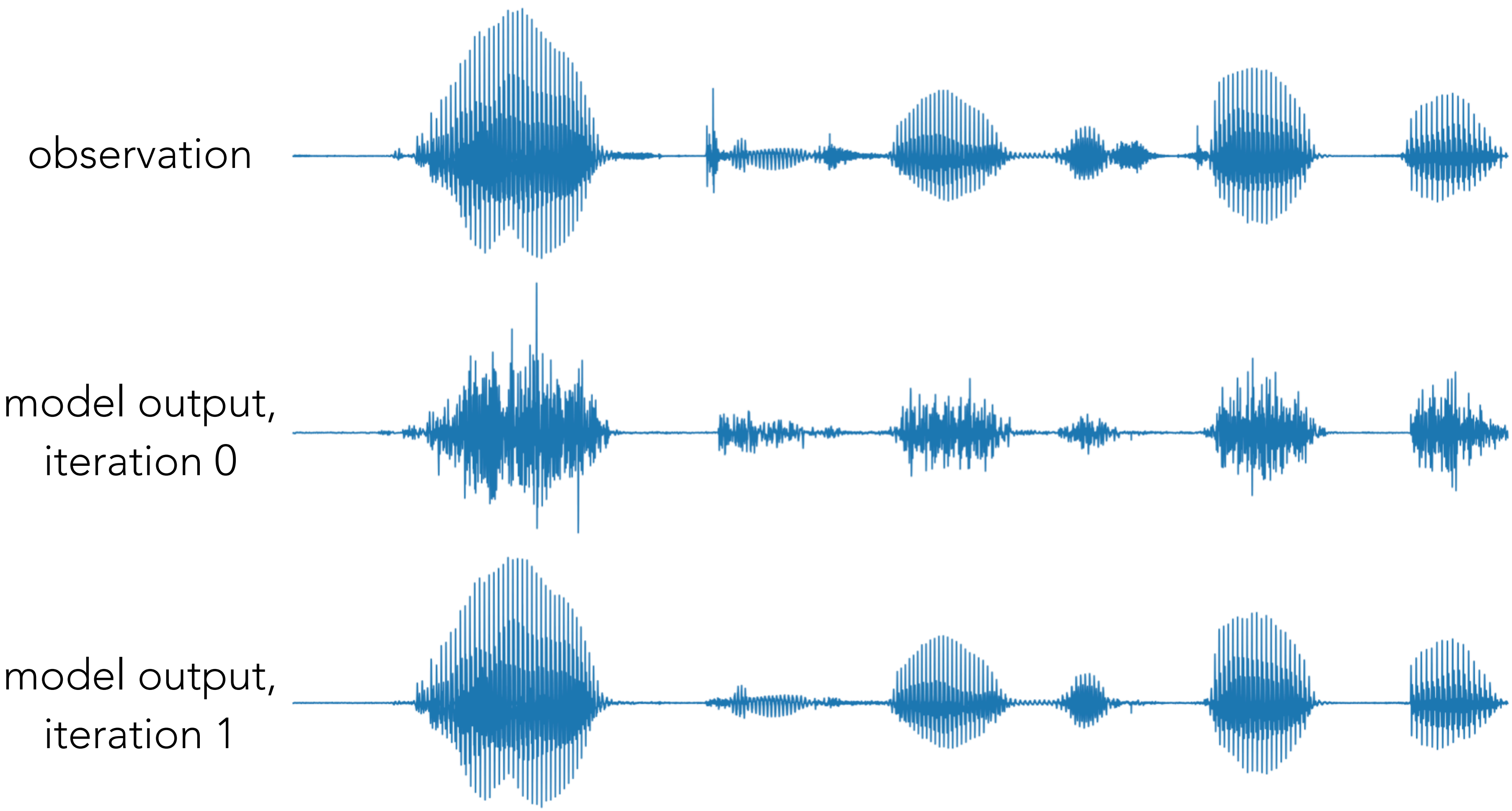


perform iterative amortized inference at each time step

Marino *et al.*, 2018b

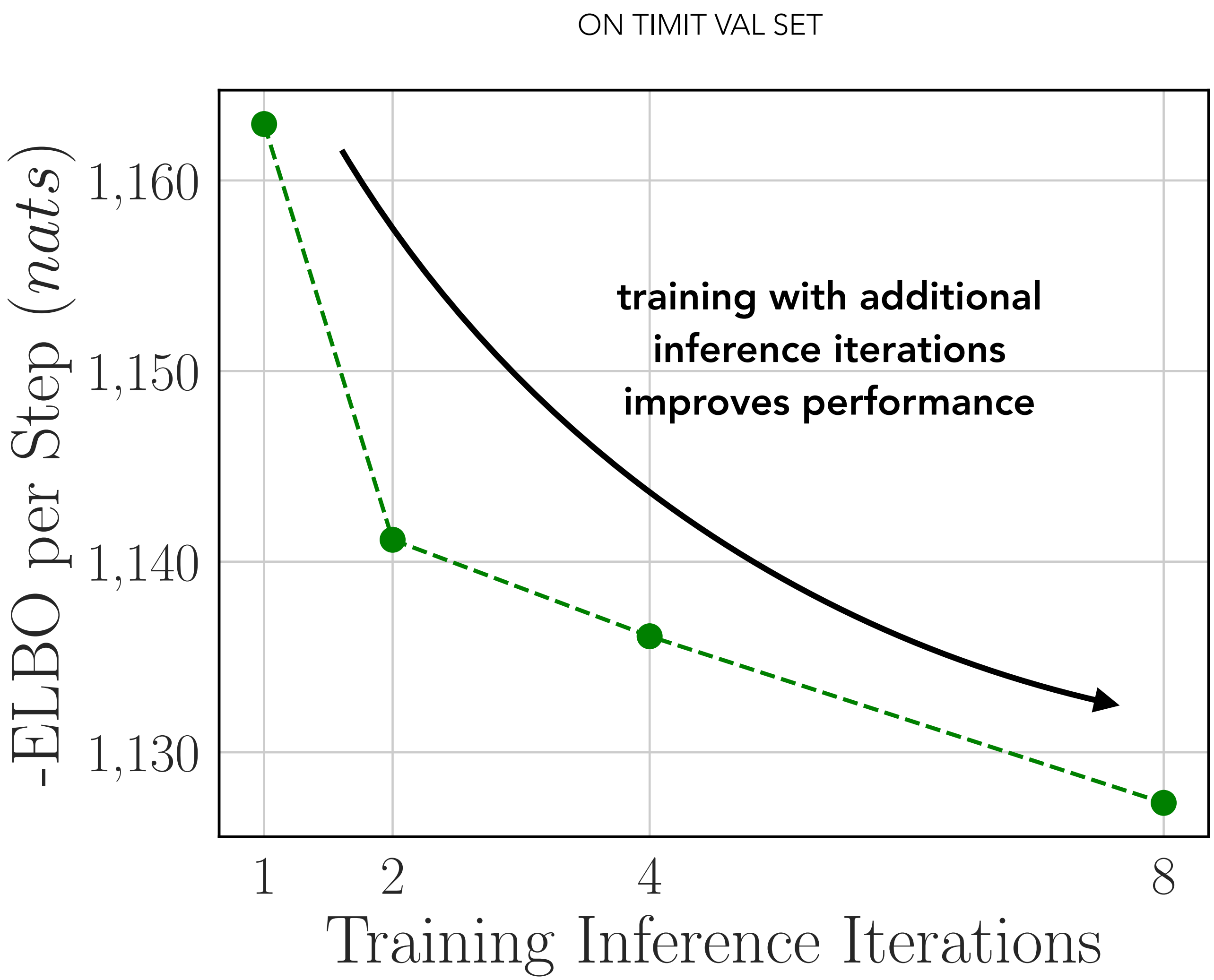
INFERENCE IMPROVEMENT

TIMIT audio waveforms



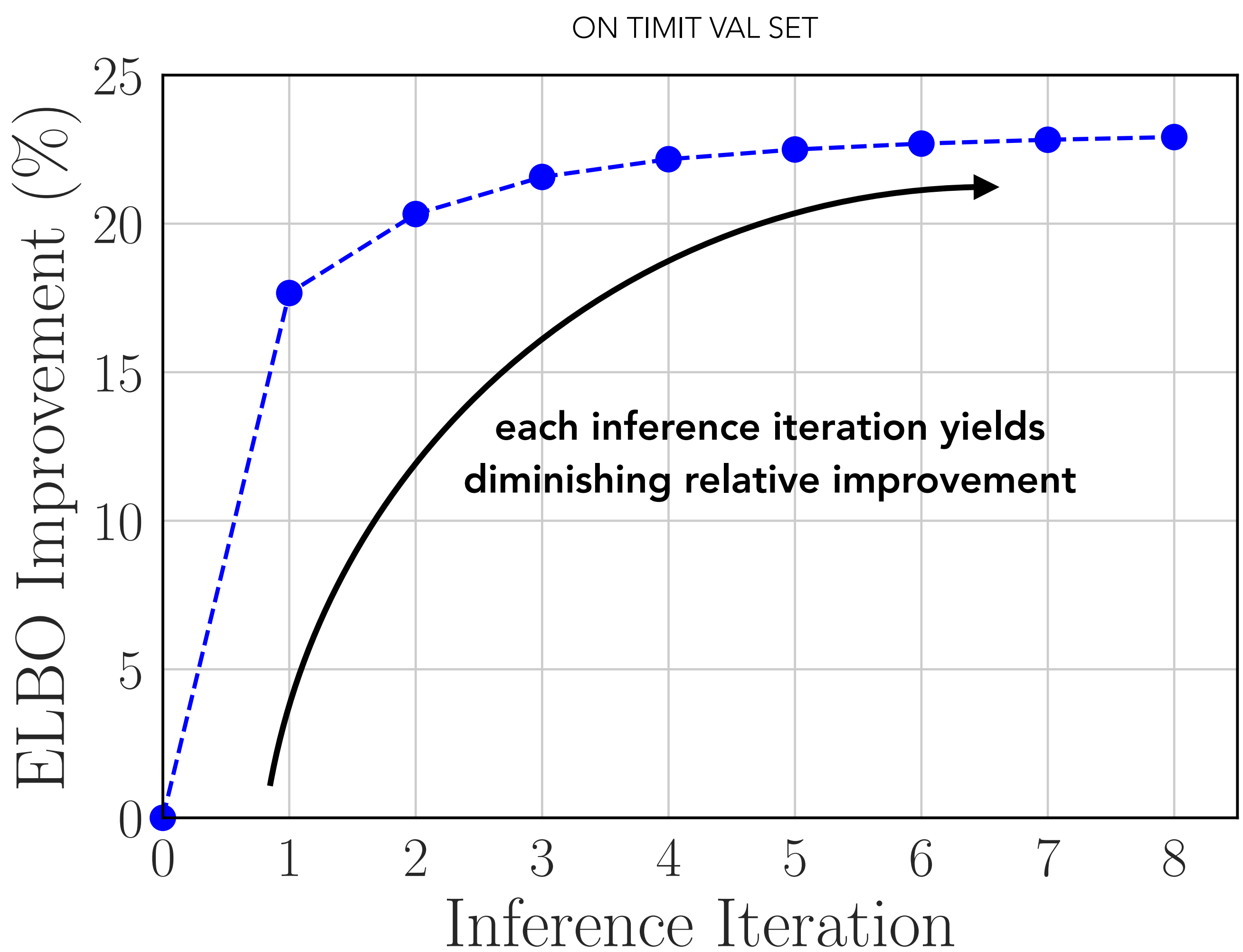
Marino *et al.*, 2018b

INFERENCE ITERATIONS



Marino *et al.*, 2018b

INFERENCE ITERATIONS



Marino *et al.*, 2018b

PERFORMANCE

<u>*one inference method*</u>, consistent improvement across models & domains

AUDIO

	TIMIT
VRNN	
baseline	1,082
AVF (1 Iter.)	1,105
AVF (2 Iter.)	**1,071**
SRNN	
baseline	1,026
AVF (1 Iter.)	**1,024**

VIDEO

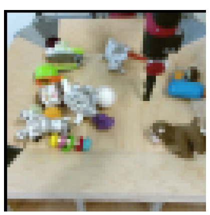
	KTH Actions
SVG	
baseline	3.69
AVF (1 Iter.)	**2.86**

MIDI MUSIC

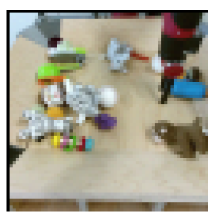
	Piano-midi.de	MuseData	JSB Chorales	Nottingham
SRNN				
baseline (Fraccaro et al., 2016)	8.20	6.28	4.74	2.94
baseline	8.19	6.27	6.92	3.19
AVF (1 Iter.)	**8.12**	**5.99**	6.97	**3.13**
AVF (5 Iter.)	–	–	**6.77**	–

Marino *et al.*, 2018b

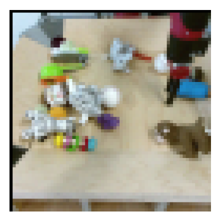
SEQUENTIAL AUTOREGRESSIVE FLOWS



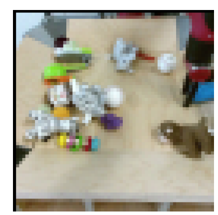
$\mathbf{x}_{t-3}$



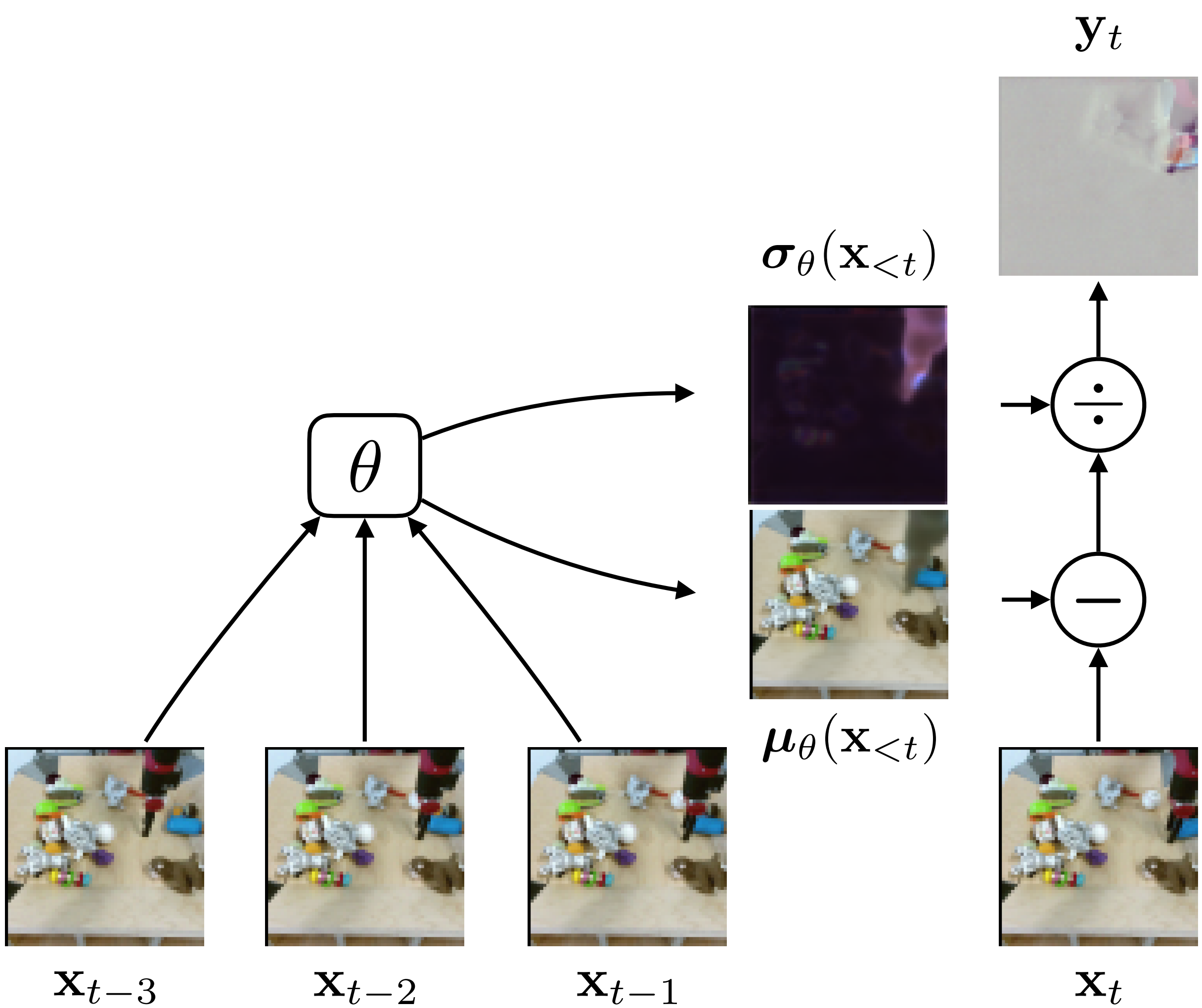
$\mathbf{x}_{t-2}$



$\mathbf{x}_{t-1}$

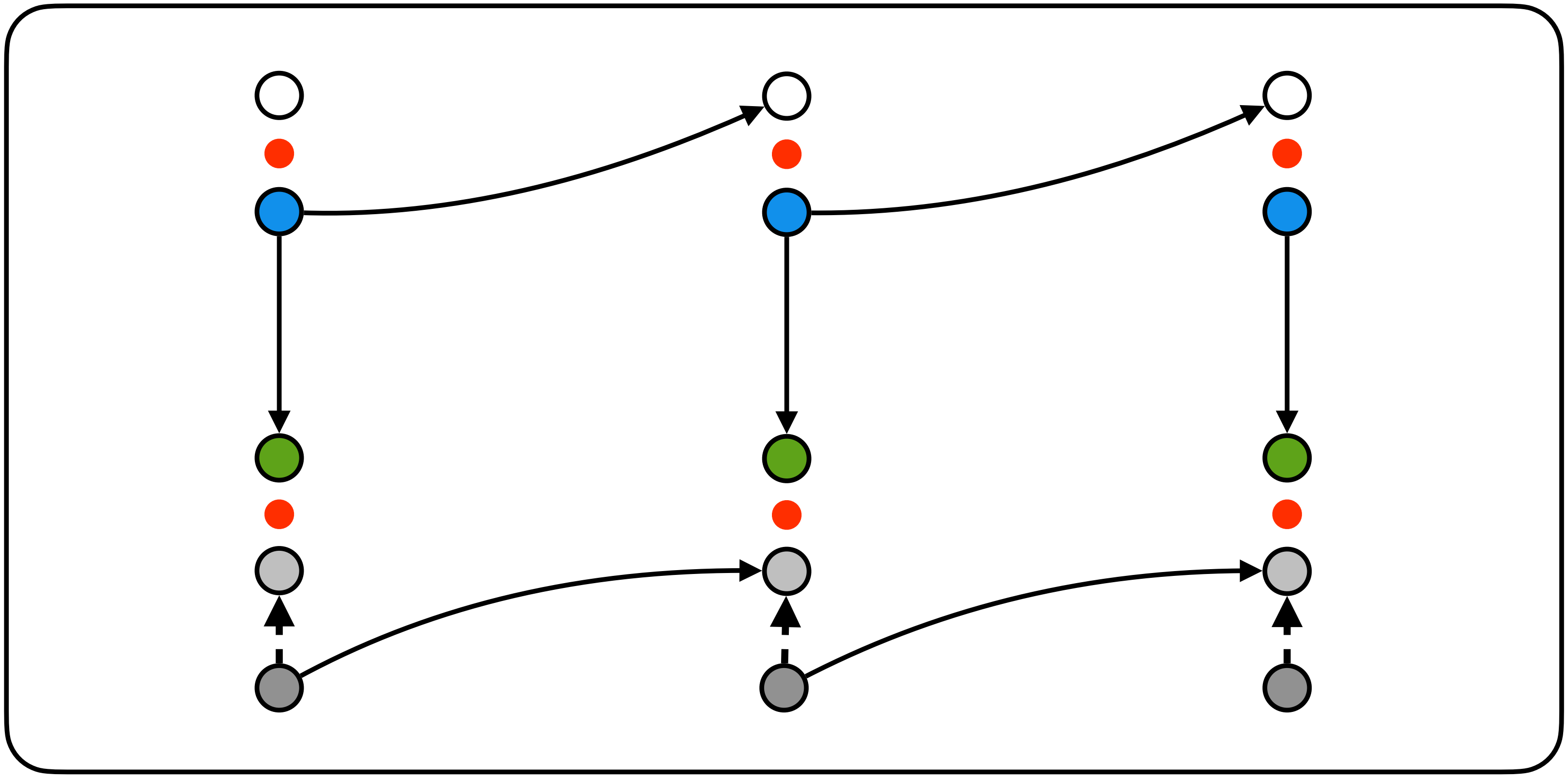


$\mathbf{x}_{t}$



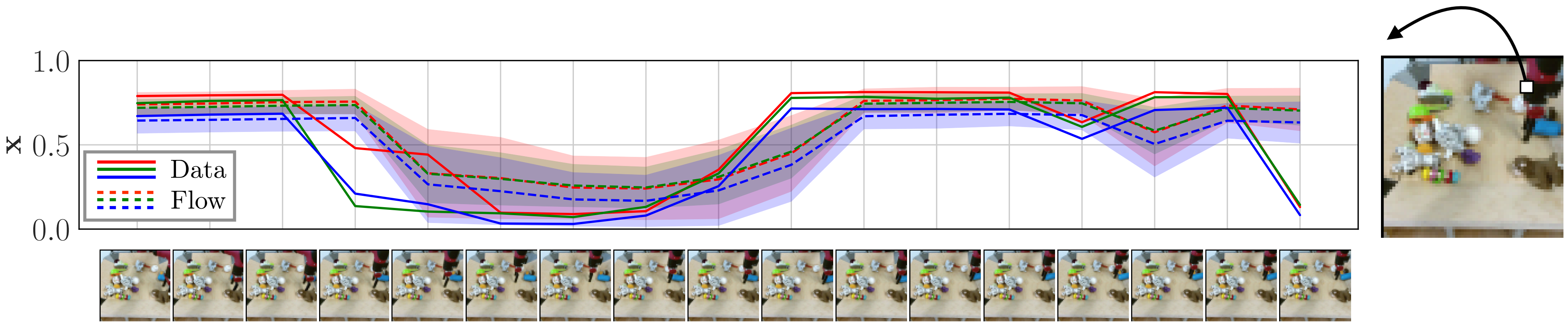
$\mathbf{y}_t$
$\boldsymbol{\sigma}_\theta(\mathbf{x}_{<t})$
θ
$\boldsymbol{\mu}_\theta(\mathbf{x}_{<t})$
$\mathbf{x}_{t-3}$
$\mathbf{x}_{t-2}$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$

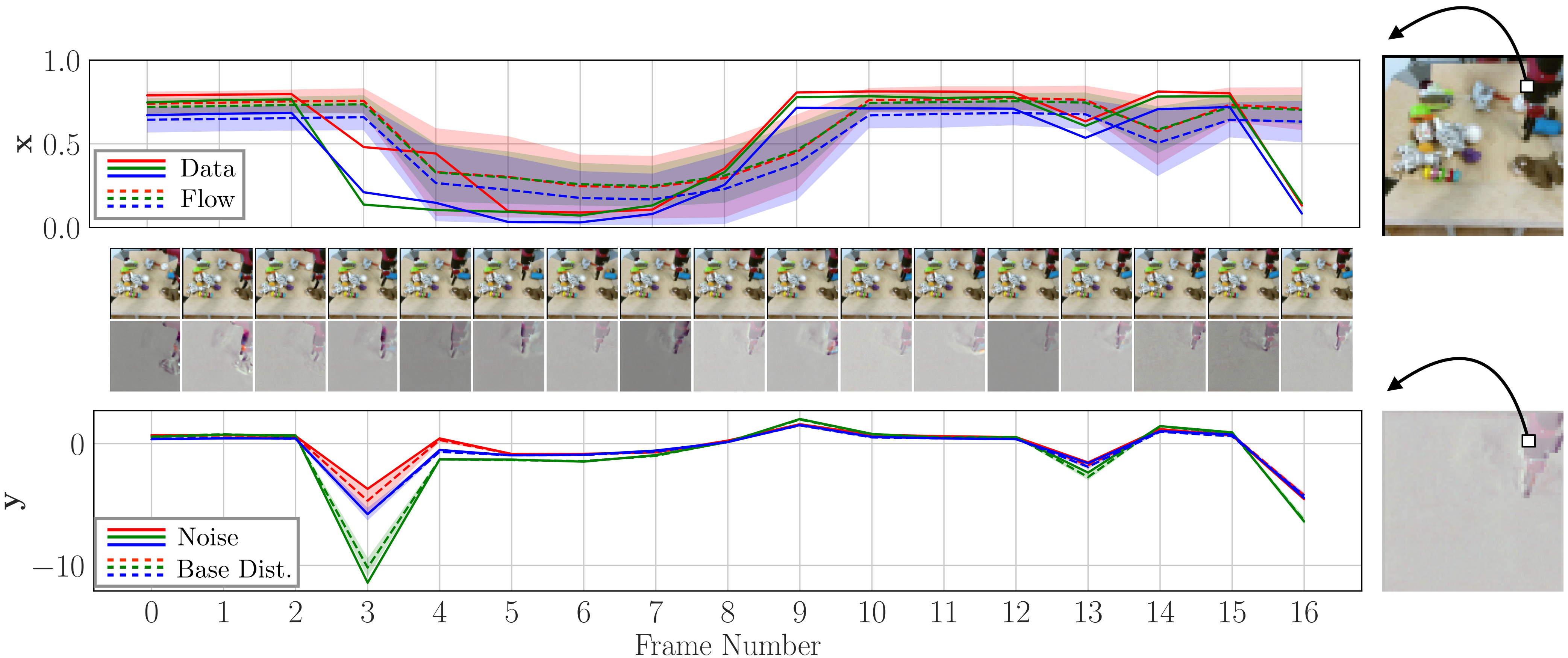
SEQUENTIAL AUTOREGRESSIVE FLOWS



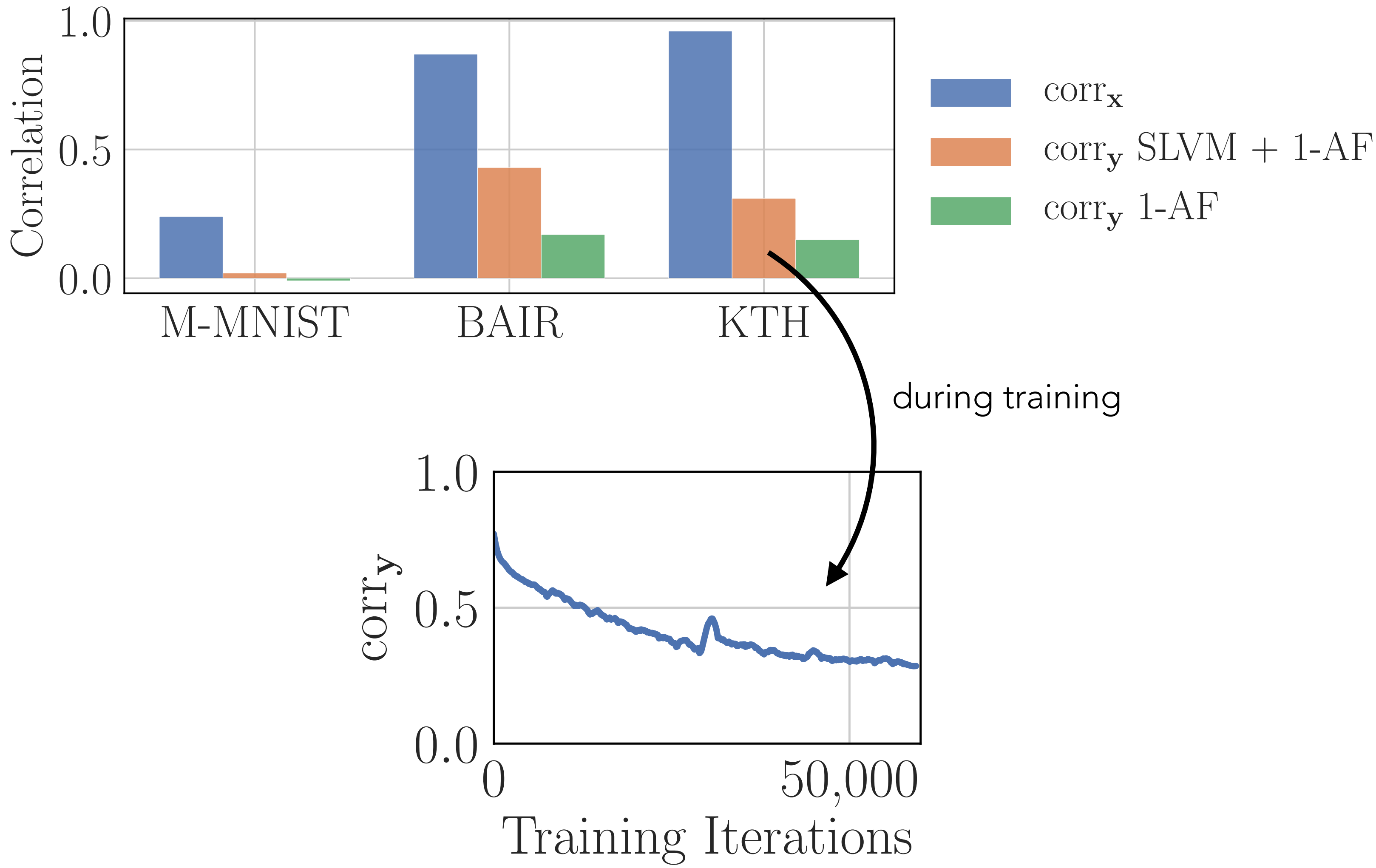
remove low-level temporal dependencies using an autoregressive flow

Marino *et al.*, 2020


1.0
0.5
0.0
x
Data
Flow


1.0
0.5
0.0
x
Data
Flow
0
−10
y
Noise
Base Dist.
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Frame Number

REDUCED TEMPORAL CORRELATION

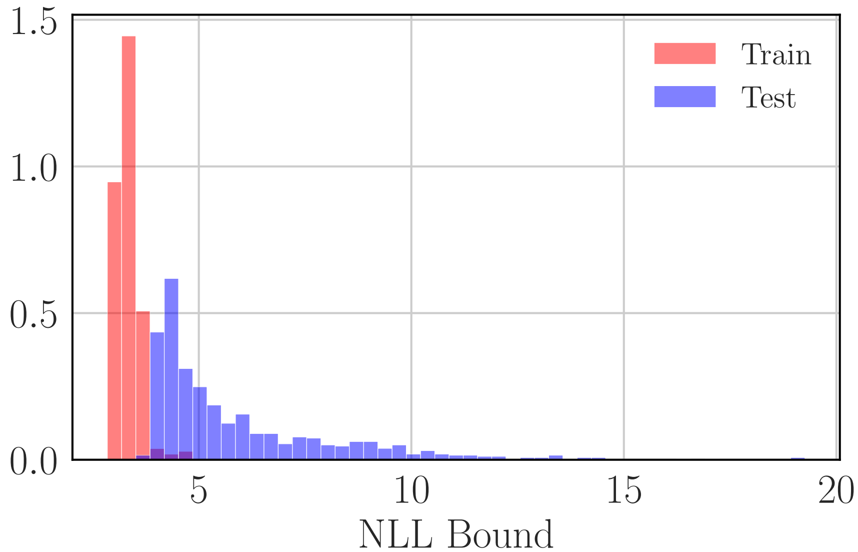


Marino *et al.*, 2020

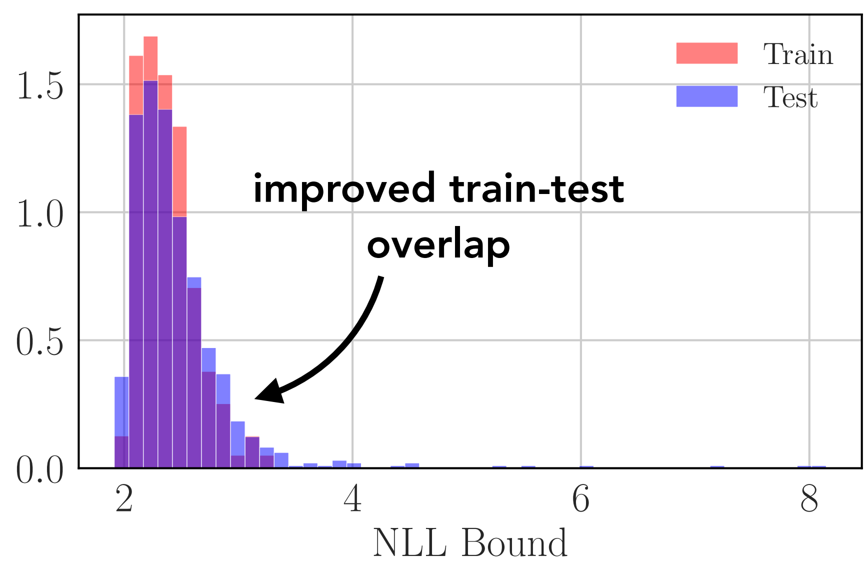
IMPROVED GENERALIZATION

KTH Actions

SLVM



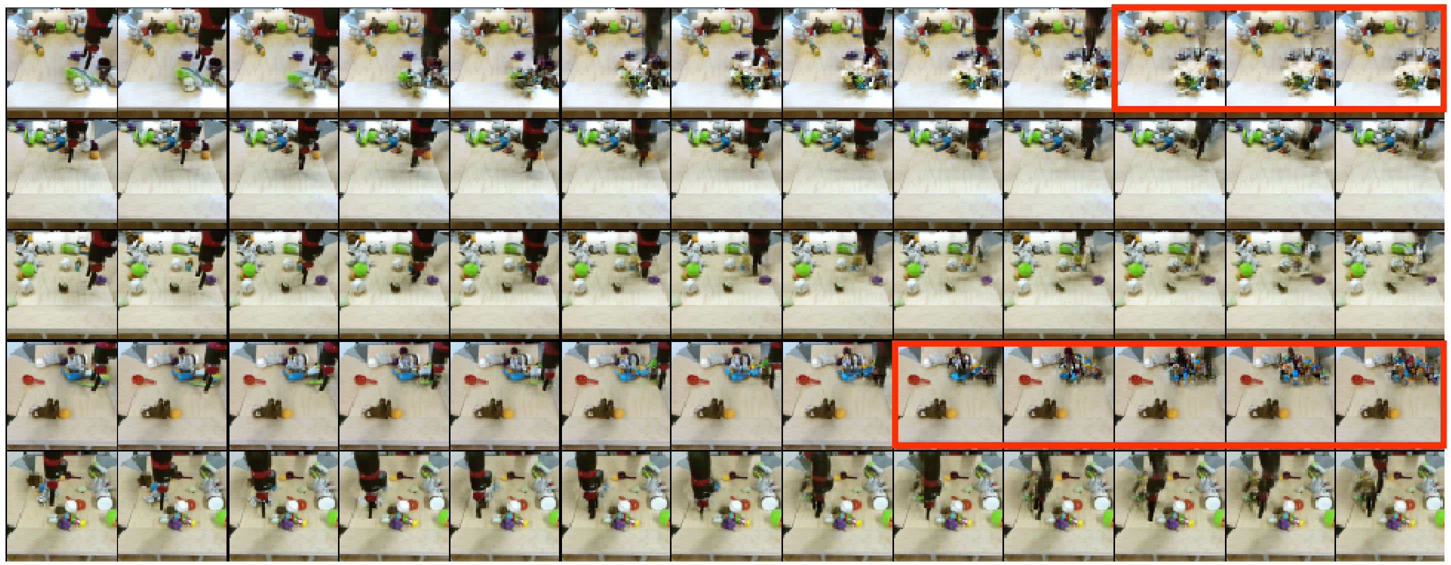
SLVM + AF



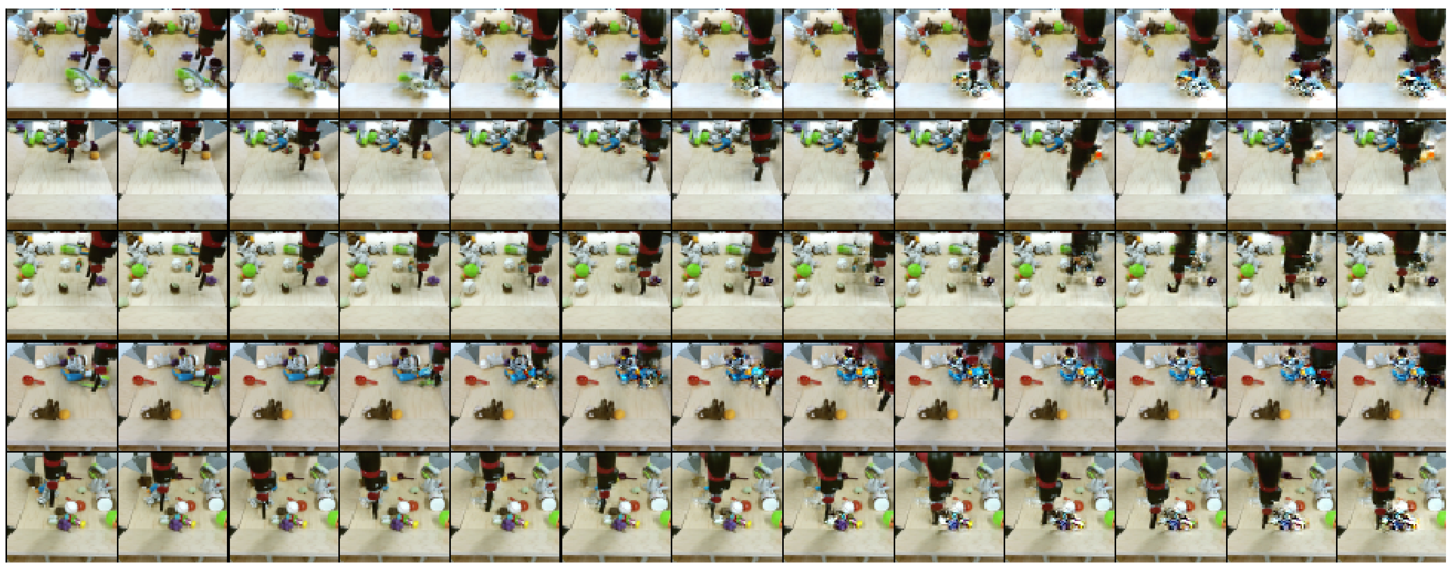
Marino *et al.*, 2020

IMPROVED SAMPLES

VideoFlow



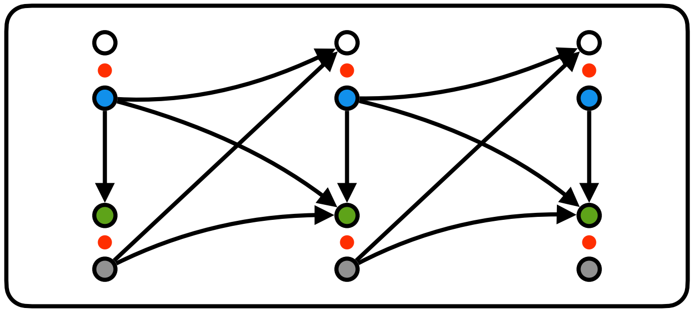
VideoFlow + AF



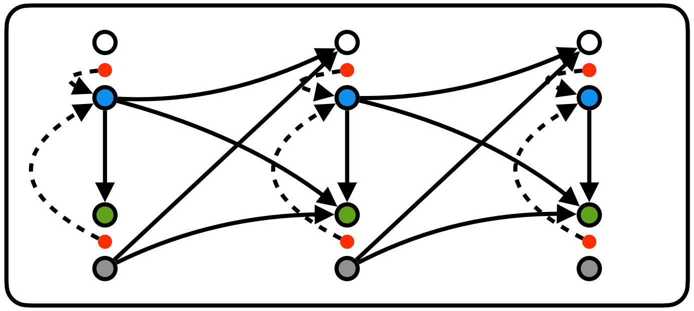
Marino *et al.*, 2020

RECAP

- sequence models



- amortized variational filtering



- sequential autoregressive flows