STOCHASTIC LATENT RESIDUAL VIDEO PREDICTION

ICML 2020

July 12th to 18th, 2020

https://sites.google.com/view/srvp/

Jean-Yves Franceschi,[1] **Edouard Delasalles**,[1]
Mickael Chen,[1] Sylvain Lamprier,[1] Patrick Gallinari[1,2]

[1]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[2]Criteo AI Lab, Paris, France

Video Prediction

Applications:

- Reinforcement Learning (Gregor et al. 2019)
- Robotics (Babaeizadeh et al. 2018)

Challenges

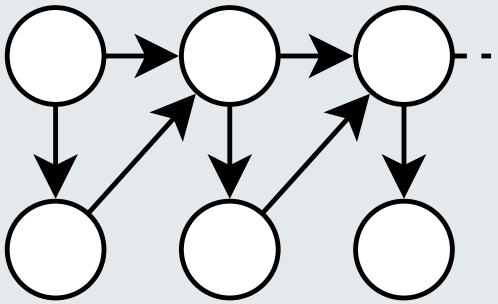
- Generation of realistic images
- Long-term prediction
- Account for uncertainty in the future

Stochastic Video Prediction Literature Taxonomy

Autoregressive Models

+ Easy to learn, powerful
– Temporal model tied to generation

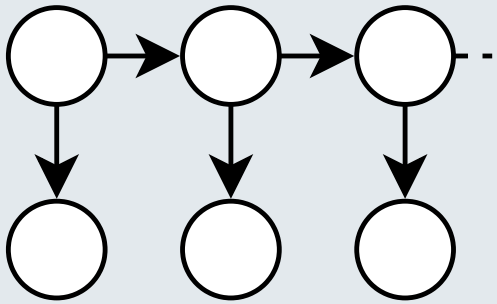
Ex.: Denton and Fergus (2018)

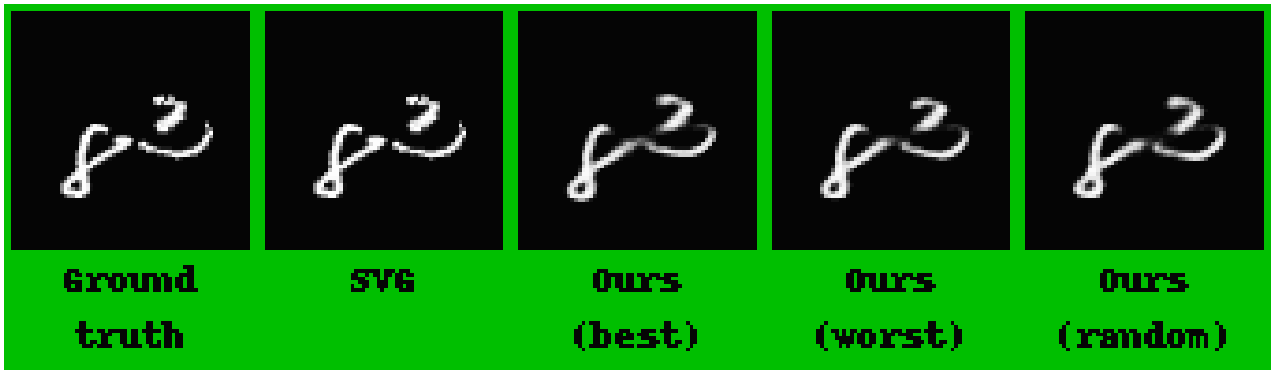


State-Space Models

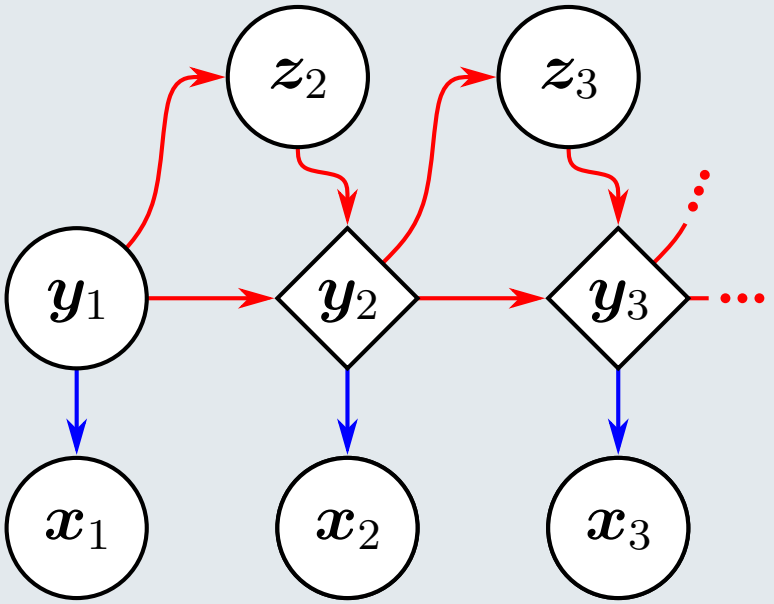
+ Decoupled dynamics and prediction, interpretable
– Harder to train

Ex.: Fraccaro et al. (2017)





Model



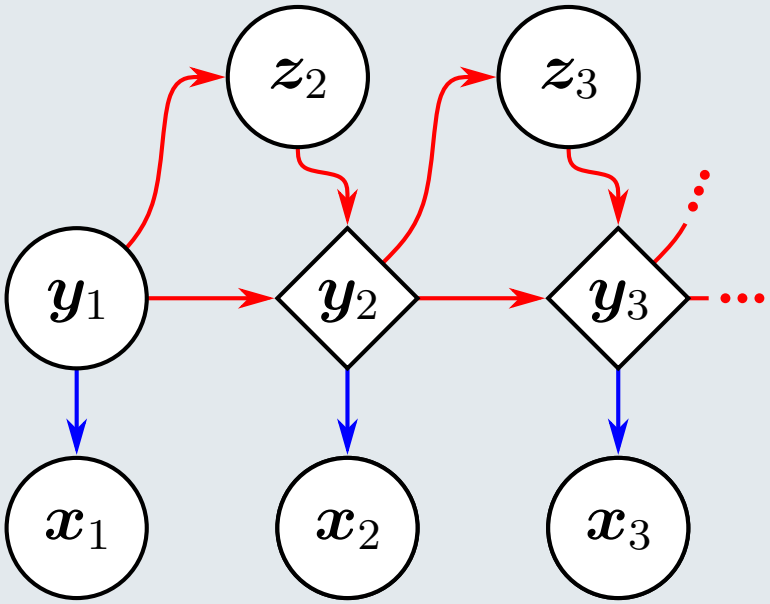
Update Rule:

$$\begin{cases} z_{t+1} \sim \mathcal{N}\big(\mu_\theta(y_t), \sigma_\theta(y_t) I\big) \\ y_{t+1} = y_t + f_\theta(y_t, z_{t+1}) \end{cases}$$

Key points:

- VAE state-space model
- Residual updates
- ODE inspiration
- Generation at arbitrary frame rates

Model



Update Rule:

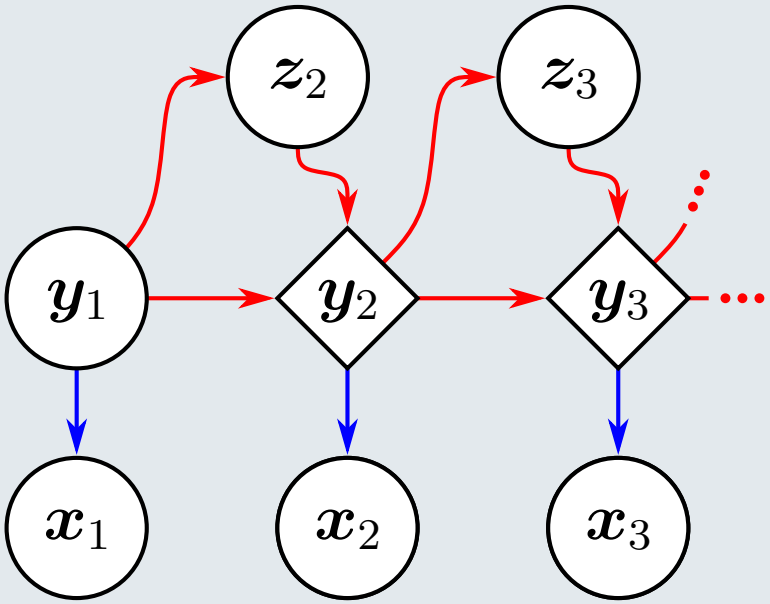
$$\begin{cases} z_{t+1} \sim \mathcal{N}\big(\mu_\theta(y_t), \sigma_\theta(y_t) I\big) \\ y_{t+1} = y_t + f_\theta(y_t, z_{t+1}) \end{cases}$$

Key points:

- VAE state-space model
- Residual updates
- ODE inspiration
- Generation at arbitrary frame rates

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f_{z_{\lfloor t \rfloor + 1}}(y)$$

Model

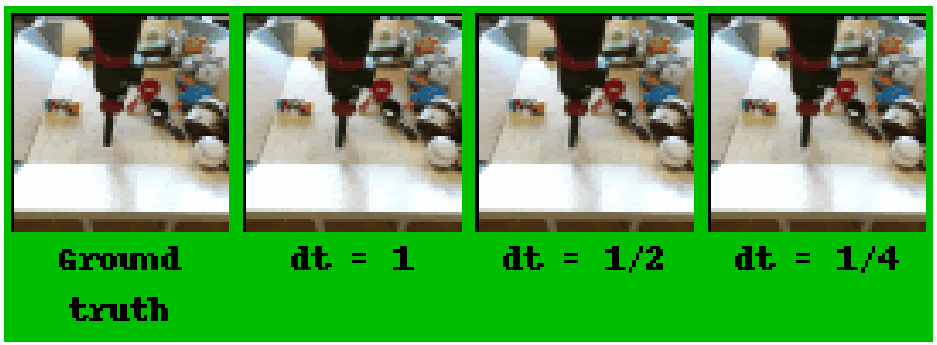


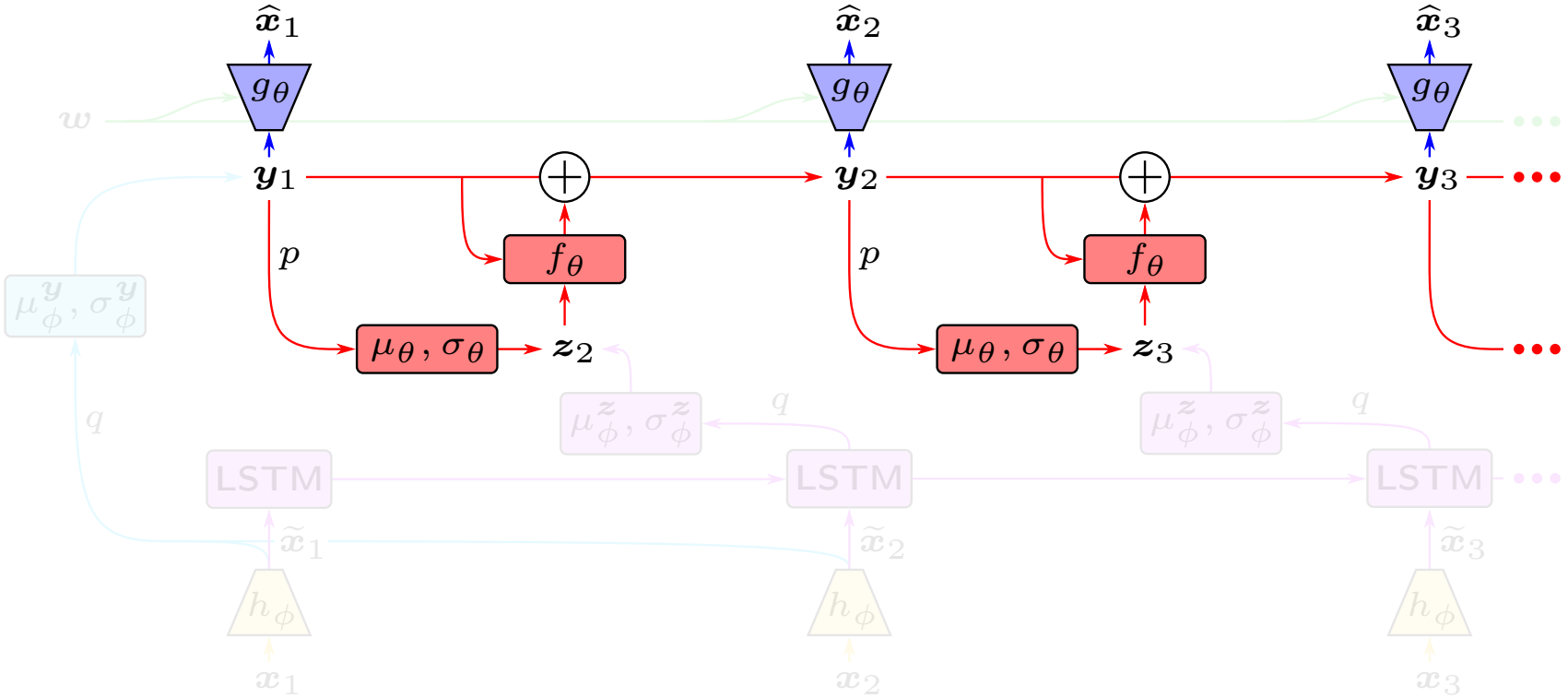
Update Rule:

$$\begin{cases} z_{t+1} \sim \mathcal{N}\big(\mu_\theta(y_t), \sigma_\theta(y_t) I\big) \\ y_{t+1} = y_t + f_\theta(y_t, z_{t+1}) \end{cases}$$

Key points:

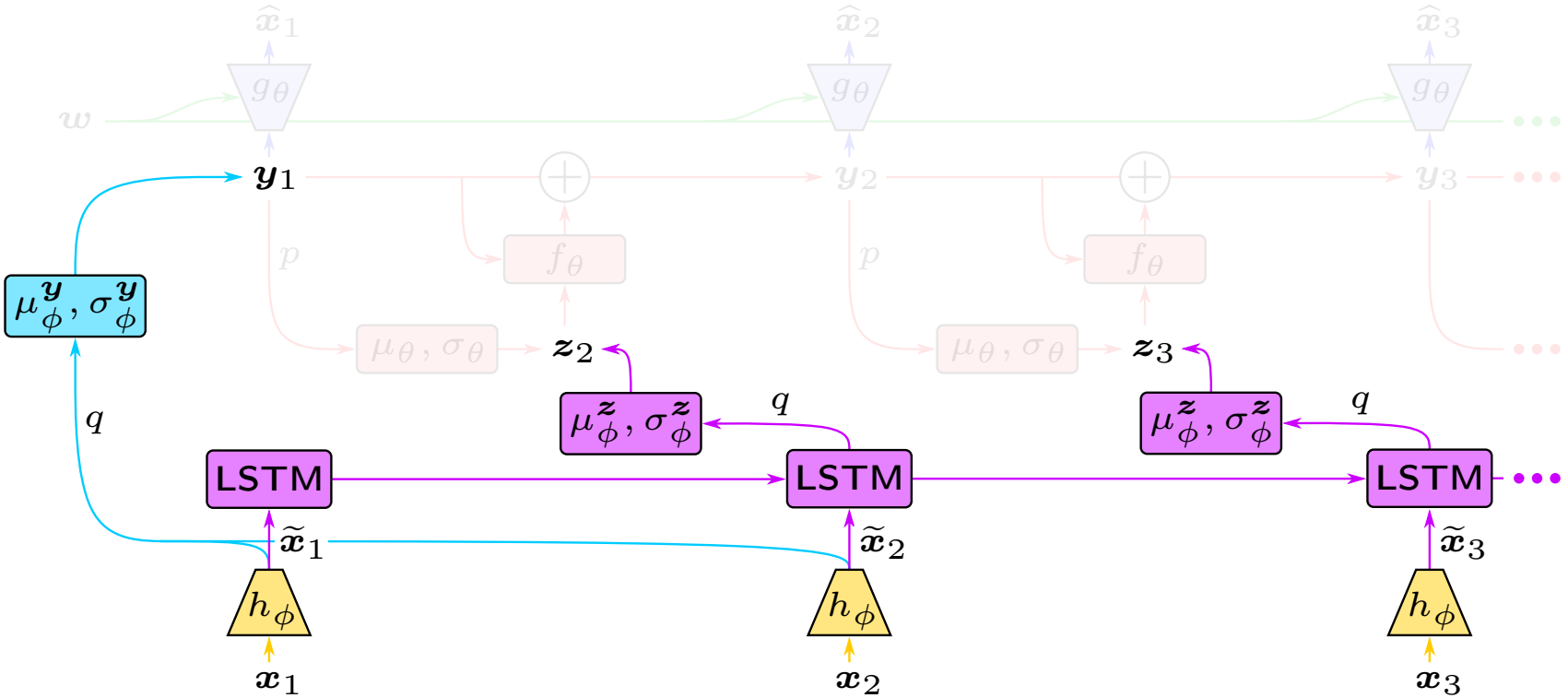
- VAE state-space model
- Residual updates
- ODE inspiration
- Generation at arbitrary frame rates





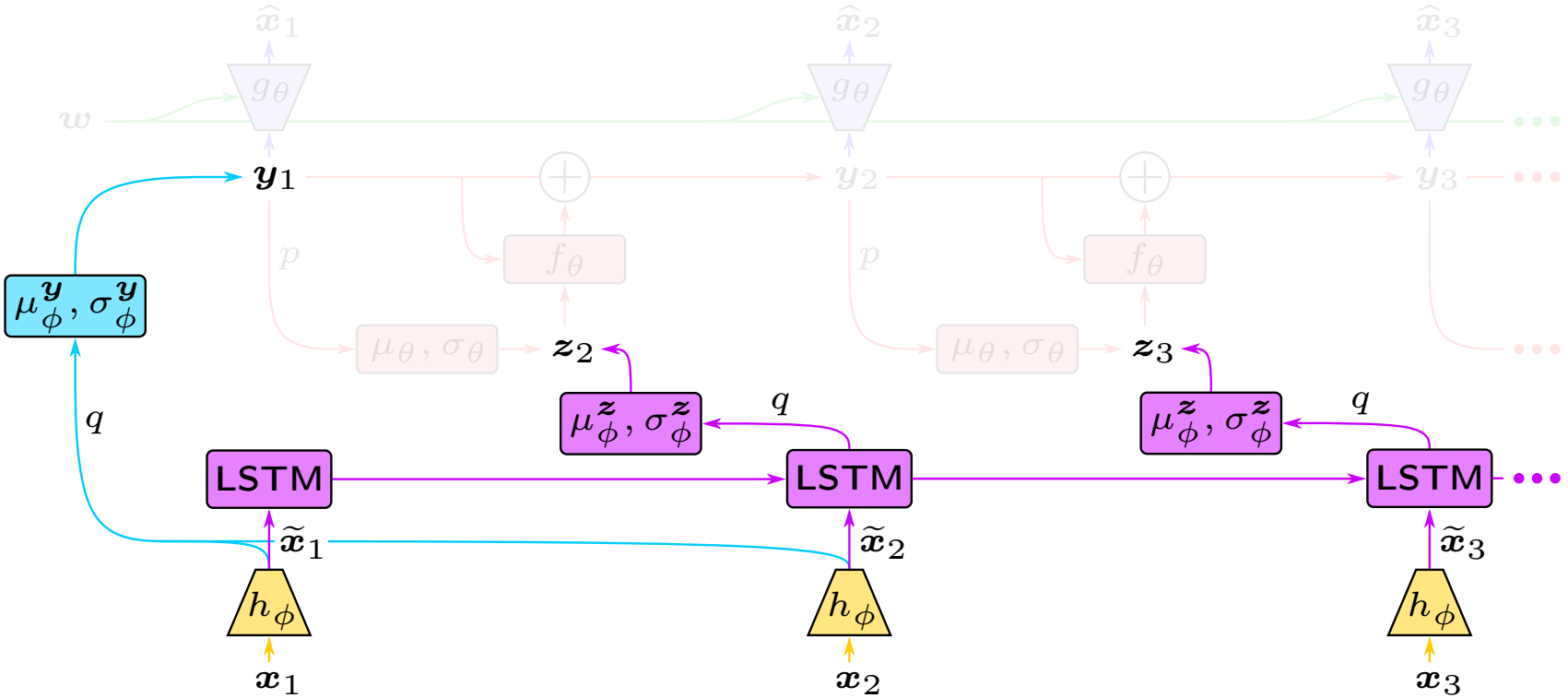
$$\begin{cases} \boldsymbol{y}_1 \sim \mathcal{N}(\mathbf{0}, I) & \text{(initial condition)} \\ \boldsymbol{z}_{t+1} \sim \mathcal{N}\big(\mu_\theta(\boldsymbol{y}_t), \sigma_\theta(\boldsymbol{y}_t) I\big) & \text{(random variable prediction)} \\ \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + f_\theta(\boldsymbol{y}_t, \boldsymbol{z}_{t+1}) & \text{(latent state prediction)} \\ \boldsymbol{x}_t \sim \mathcal{N}\big(g_\theta(\boldsymbol{y}_t), \nu I\big) & \text{(decoding)} \end{cases}$$

Inference



$$q(\boldsymbol{z}_{2:T}, \boldsymbol{y}_1 \mid \boldsymbol{x}_{1:T}) = \underbrace{q(\boldsymbol{y}_1 \mid \boldsymbol{x}_{1:k})}_{\text{Init. Cond.}} \prod_{t=2}^{T} \underbrace{q(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})}_{\text{LSTM}}$$

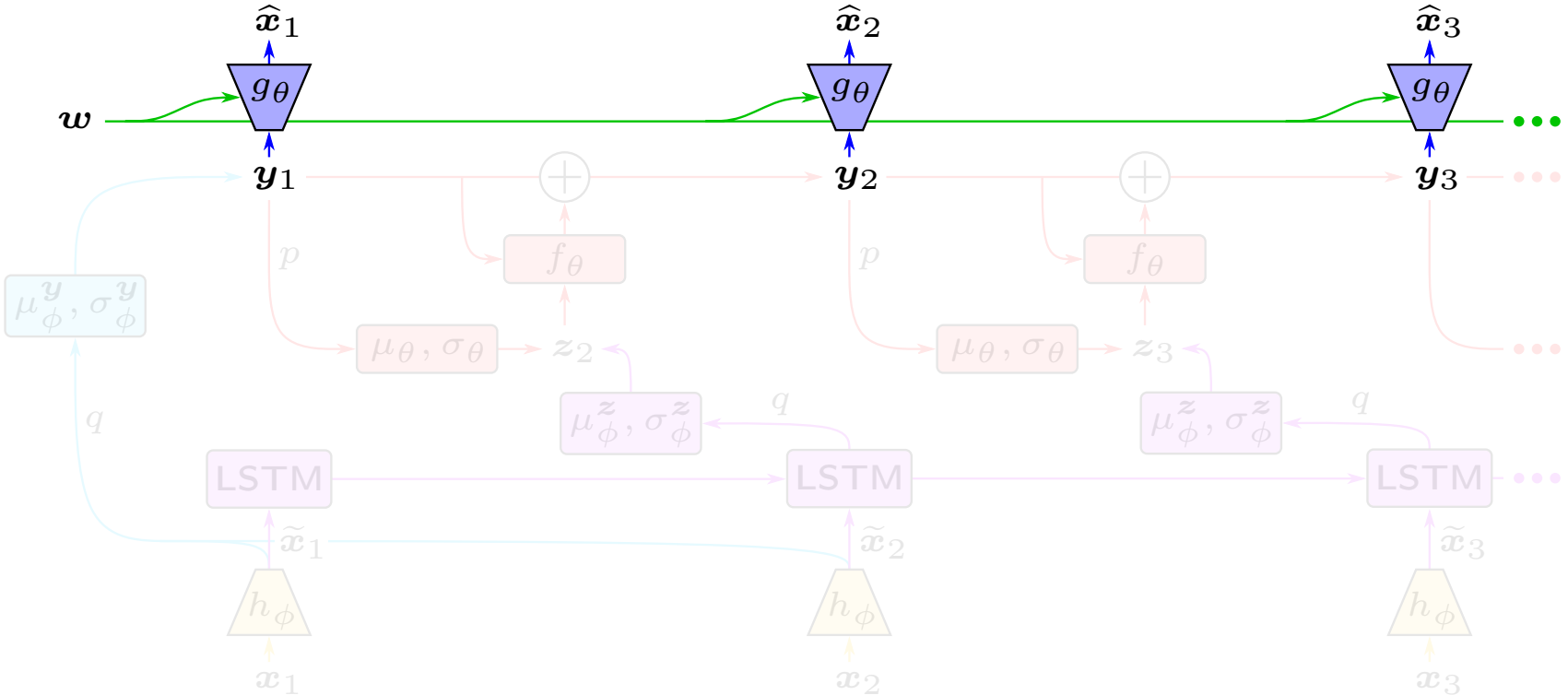
Inference



$$q(\boldsymbol{z}_{2:T}, \boldsymbol{y}_1 \mid \boldsymbol{x}_{1:T}) = \underbrace{q(\boldsymbol{y}_1 \mid \boldsymbol{x}_{1:k})}_{\text{Init. Cond.}} \prod_{t=2}^{T} \underbrace{q(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})}_{\text{LSTM}}$$

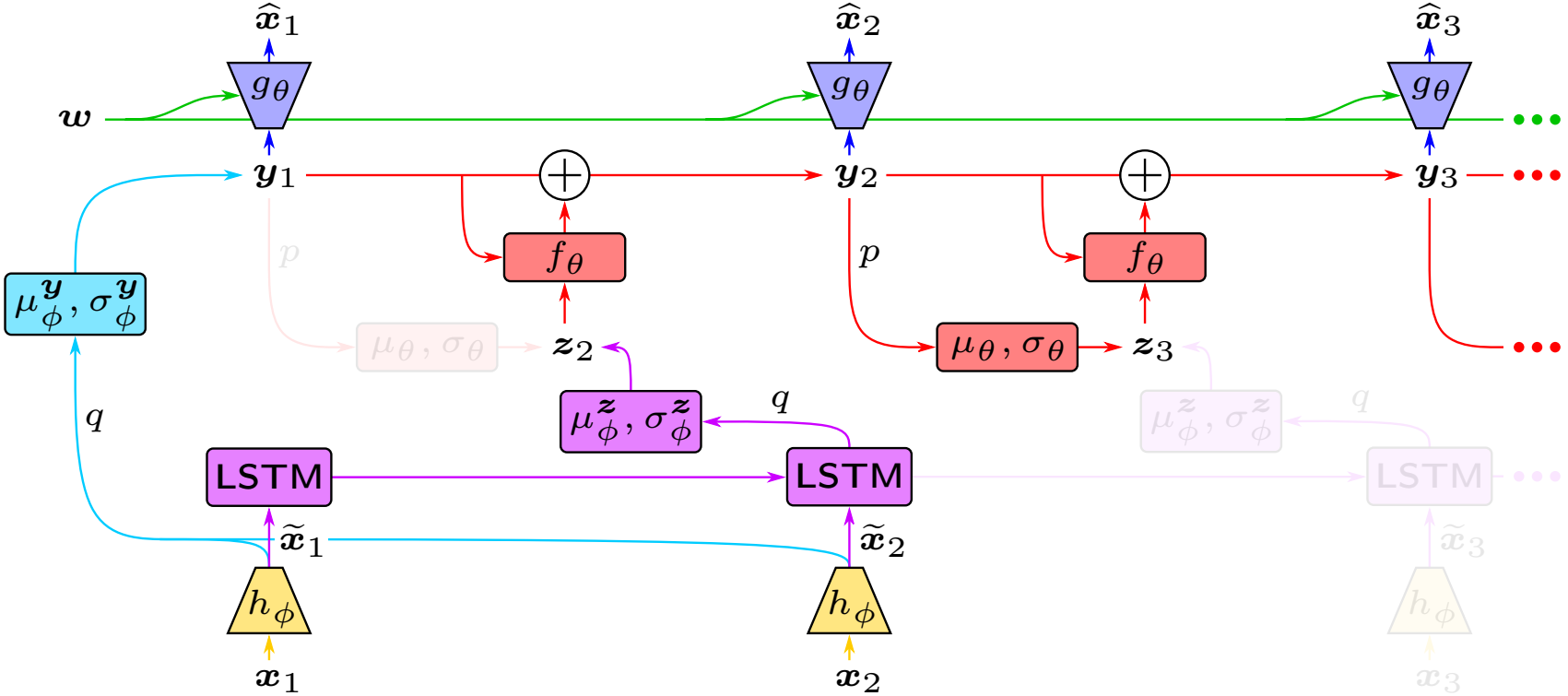
Training done using variational inference within an ELBO objective.

Content Variable w



- Store static information (e.g., background and object shapes)
- Computed from randomly sampled frames $\rightarrow$ temporal invariance
- Skip connections between encoder and decoder

Evaluation

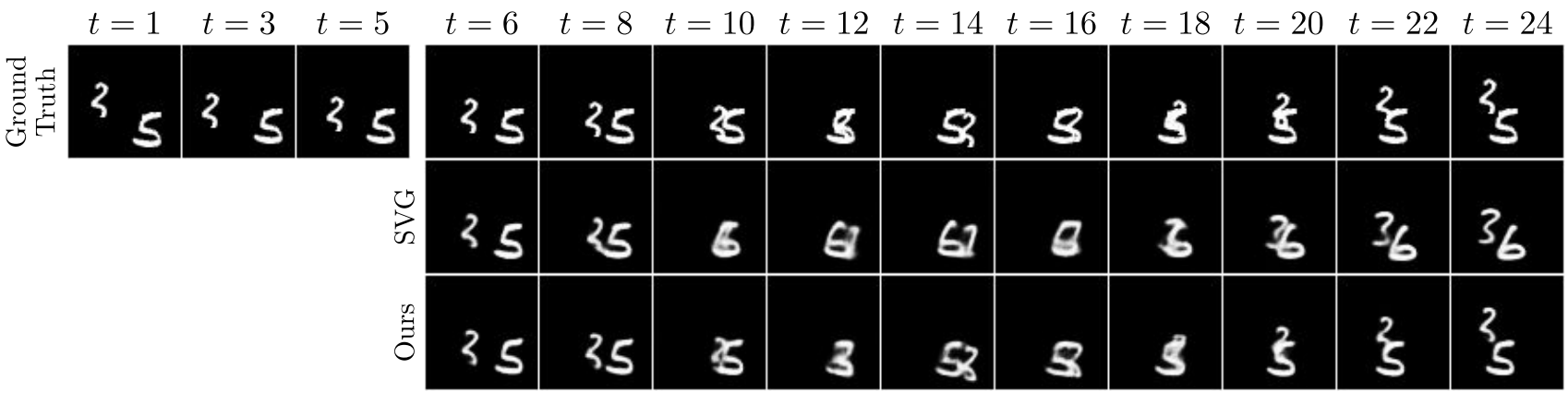


- Conditioning frames are used to infer dynamic variables
- Prediction follows using the forward model

Experimental Results: Moving MNIST

Models	Stochastic		Deterministic	
	PSNR	SSIM	PSNR	SSIM
SVG	14.50	0.7090	12.85	0.6185
Ours	**16.93**	**0.7799**	**18.25**	**0.8300**
Ours - GRU	15.80	0.7464	13.17	0.6237
Ours - MLP	16.55	0.7694	16.70	0.7876
Ours - w/o z	—	—	14.99	0.4757

SVG: Denton and Fergus (2018)

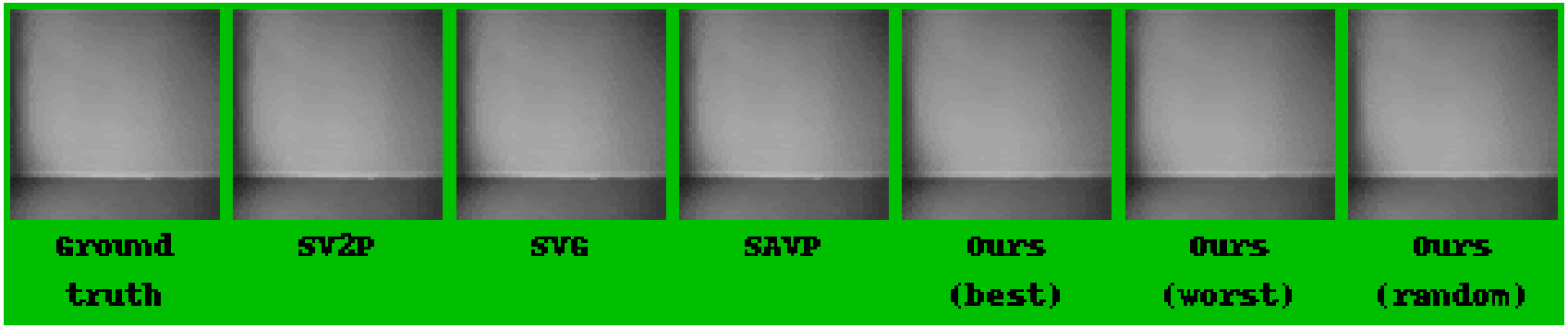


Experimental Results: KTH, Human3.6M and BAIR

Metric	Dataset	SV2P	SAVP	SVG	SVRNN	Ours
FVD (↓)	KTH	636	374	377	—	**222**
	H3.6M	—	—	—	556	**416**
	BAIR	965	**152**	255	—	163
LPIPS (↓)	KTH	0.2049	0.1120	0.0923	—	**0.0736**
	H3.6M	—	—	—	0.0557	**0.0509**
	BAIR	0.0912	0.0634	0.0609	—	**0.0574**
PSNR (↑)	KTH	28.19	26.51	28.06	—	**29.69**
	H3.6M	—	—	—	24.46	**25.30**
	BAIR	**20.39**	18.44	18.95	—	19.59

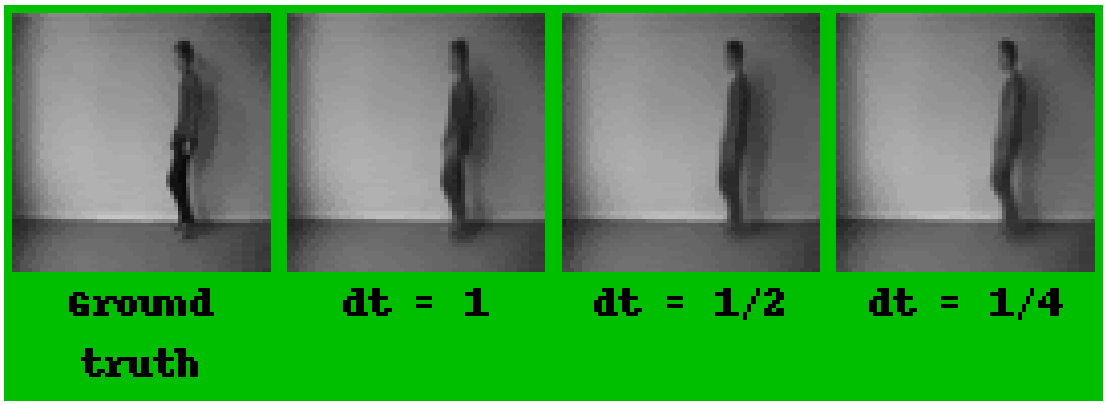
SV2P: Finn, Goodfellow, and Levine (2016), SAVP: Babaeizadeh et al. (2018), SVRNN: Minderer et al. (2019)

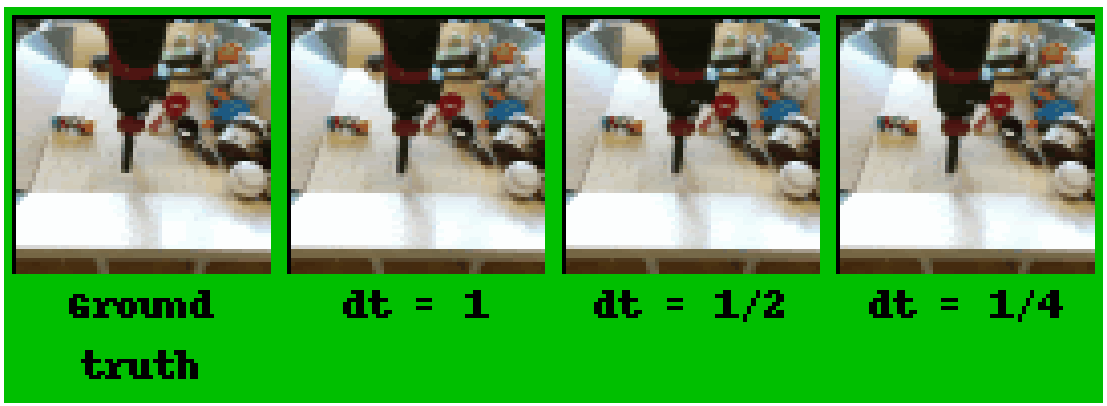
Experimental Results: Samples



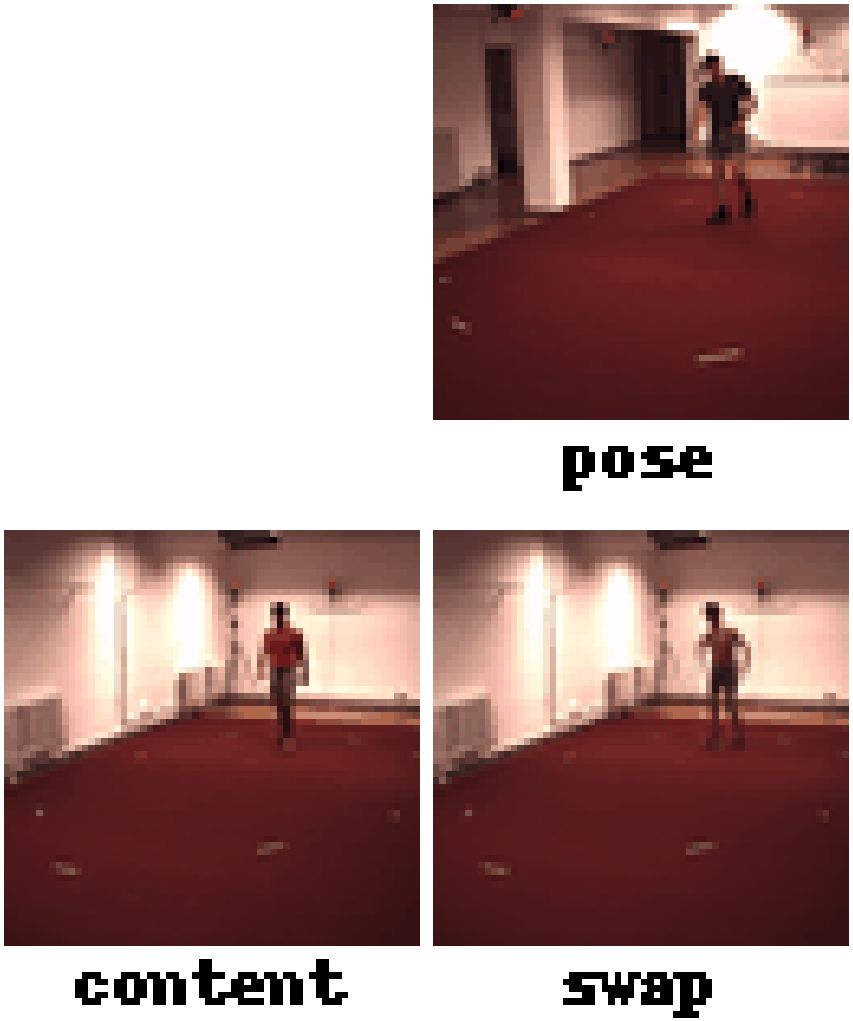
Using the Euler approximation scheme:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = f_{\boldsymbol{z}_{\lfloor t \rfloor + 1}}(\boldsymbol{y}) \quad \Rightarrow \quad \boldsymbol{y}_{t+\Delta t} = \boldsymbol{y}_t + \Delta t \cdot f_\theta\Big(\boldsymbol{y}_t, \boldsymbol{z}_{\lfloor t \rfloor + 1}\Big)$$





Experimental Results: Content Swap



- State-space video prediction model
- Based on performant residual updates
- With near-continuous dynamics
- Perspectives:
  - Additional inductive biases
  - Scaled models
  - Uses in other domains

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = f_{\boldsymbol{z}_{\lfloor t \rfloor + 1}}(\boldsymbol{y})$$

Conclusion

- State-space video prediction model
- Based on performant residual updates
- With near-continuous dynamics
- Perspectives:
  - Additional inductive biases
  - Scaled models
  - Uses in other domains

Animated samples, code and pretrained models

`https://sites.google.com/view/srvp/`

References

- Babaeizadeh, Mohammad et al. (2018). "Stochastic Variational Video Prediction". In: *International Conference on Learning Representations*.
- Denton, Emily and Rob Fergus (July 2018). "Stochastic Video Generation with a Learned Prior". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm, Sweden: PMLR, pp. 1174–1183.
- Finn, Chelsea, Ian Goodfellow, and Sergey Levine (2016). "Unsupervised Learning for Physical Interaction through Video Prediction". In: *Advances in Neural Information Processing Systems 29*. Ed. by Daniel D. Lee et al. Curran Associates, Inc., pp. 64–72.
- Fraccaro, Marco et al. (2017). "A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning". In: *Advances in Neural Information Processing Systems 30*. Ed. by Isabelle Guyon et al. Curran Associates, Inc., pp. 3601–3610.
- Gregor, Karol et al. (2019). "Temporal Difference Variational Auto-Encoder". In: *International Conference on Learning Representations*.
- Minderer, Matthias et al. (2019). "Unsupervised learning of object structure and dynamics from videos". In: *Advances in Neural Information Processing Systems 32*. Ed. by Hanna Wallach et al. Curran Associates, Inc., pp. 92–102.