SCIENCES SORBONNE UNIVERSITÉ

STOCHASTIC LATENT RESIDUAL VIDEO PREDICTION

LOCUST @ Deezer

Tuesday 8th October, 2019

Jean-Yves Franceschi*, Edouard Delasalles*,
Mickael Chen, Sylvain Lamprier, Patrick Gallinari

LIP6

MLIA Machine Learning & Deep Learning for Information Access

Video Prediction

Video Prediction

Given a few conditioning frames, learning the distribution of future frames.

Challenges

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

Applications

- Reinforcement Learning [Gre+19]
- Robotics [Bab+18]

) Challenges the ability of a model to capture visual and dynamic representations of the world

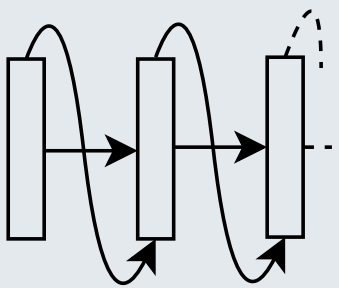
Related Work

Image Auto-Regressive Approaches [Bab+18] [DF18] [Lee+18]

Uses previously generated frames to predict next one

(+) Easy to learn, based on powerful LSTMs, work with high-definition images

( ) High computational cost, prediction tied to image generation

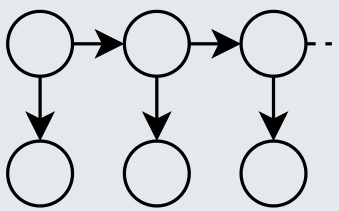


State-Space Models [Fra+16] [Kar+17] [Fra+17] [KSS17]

Prediction dynamics is fully latent. Latent states are independently decoded into frames.

(+) Fast latent prediction, not all frames have to be decoded, interpretable latent space

( ) Harder to train, only proposed on low dimensional data



Method

Our Approach

Latent Residual Dynamic Model

- Deterministic latent state y_t
- Stochastic auxiliary variable z_t
- Residual update of y_t
  - *!* Integration of ODEs in neural network architectures [Che+18]
  - *!* Near-continuous dynamics
  - *!* Generation at arbitrary frame rates
- End-to-end training with variational inference (VAE)

Content Variable

- Store "static" information (background, object shapes, etc...)
- Inferred from randomly sampled frame *!* temporal invariance
- Skip-connections between encoder and decoder

Latent Residual Temporal Model

- x_t: t-th frame
- y_t: latent space corresponding to x_t
- z_t: random variable encapsulating stochasticity at time t
- f : residual function
- g : decoder

$$
\begin{cases}
y_1 \quad N(0; I) & \text{(initial condition)} \\
z_{t+1} \quad N \quad (y_t); \quad (y_t) I & \text{(random variable prediction)} \\
y_{t+1} = y_t + f\ (y_t; z_{t+1}) & \text{(latent state prediction)} \\
x_t \quad N\ g\ (y_t); \quad I & \text{(decoding)}
\end{cases}
$$

Latent Residual Temporal Model

- x_t: t-th frame
- y_t: latent space corresponding to x_t
- z_t: random variable encapsulating stochasticity at time t
- f : residual function
- g : decoder

$$
\begin{cases}
y_1 \quad N(0; I) & \text{(initial condition)} \\
z_{t+1} \quad N \quad (y_t); \quad (y_t) I & \text{(random variable prediction)} \\
y_{t+1} = y_t + f\ (y_t; z_{t+1}) & \text{(latent state prediction)} \\
x_t \quad N\ g\ (y_t); \ I & \text{(decoding)}
\end{cases}
$$

Extension Leveraging the Euler Analogy

$$y_{t+\ \ t} = y_t + \quad t \quad f \quad y_t; z_{btc+1} \ :$$

Visual Representation

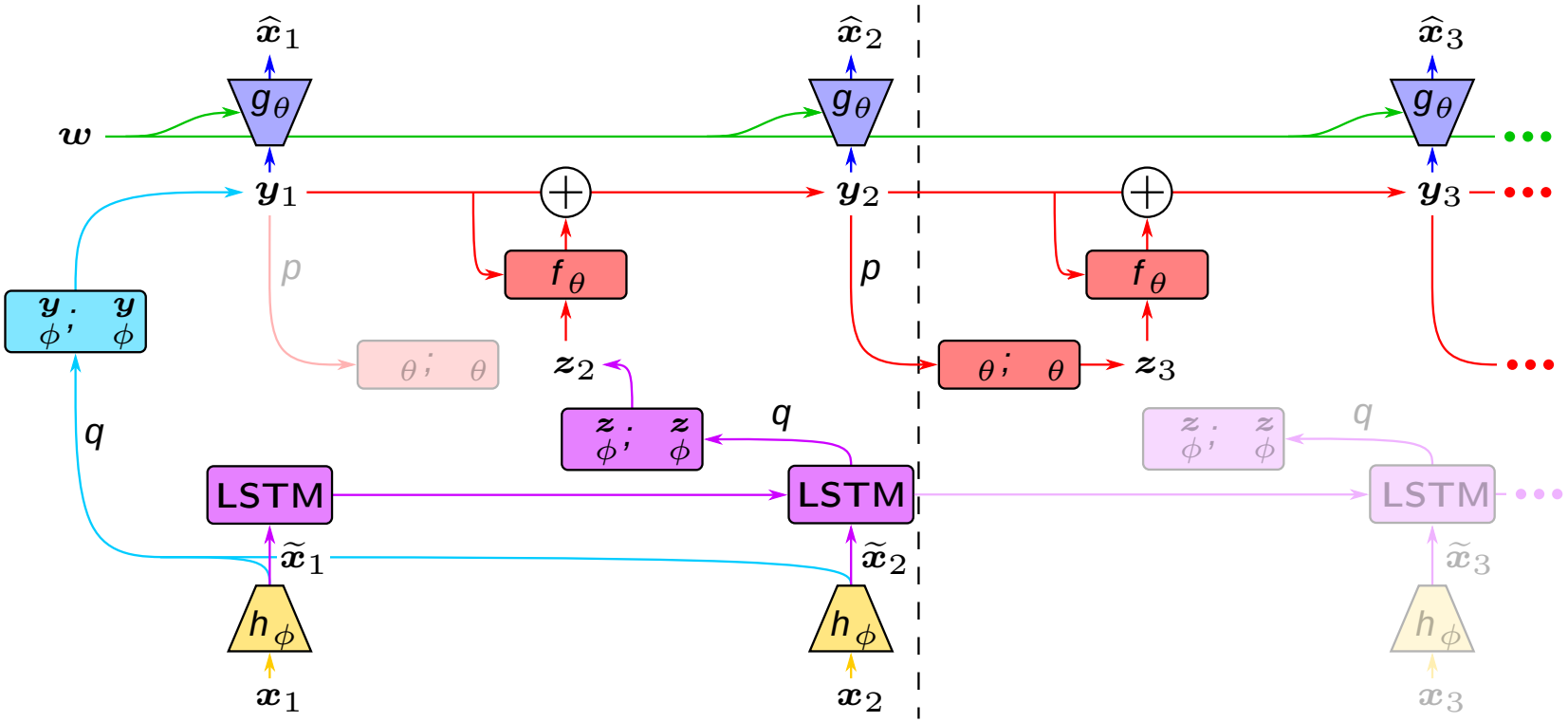
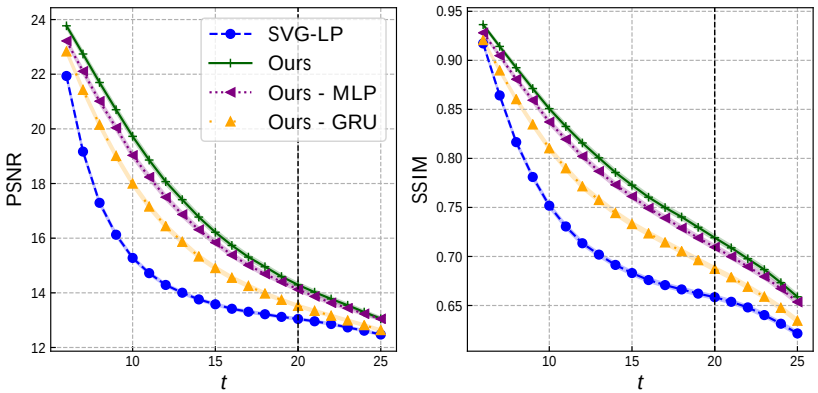


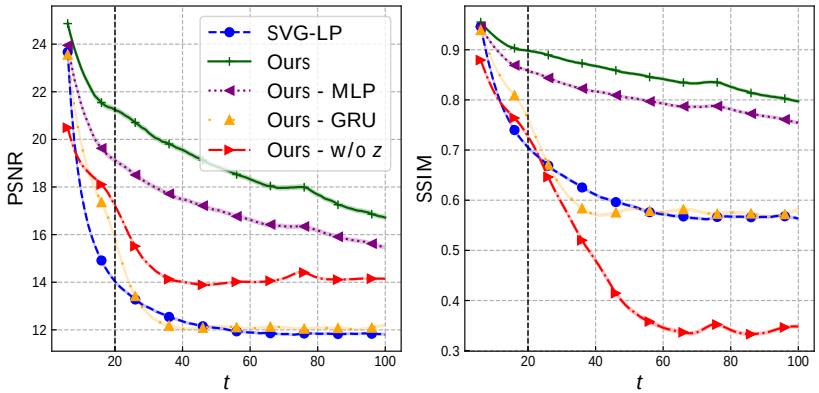
Figure 1: Model ( ) and inference (, LSTM) architecture on a test sequence.

Experiments

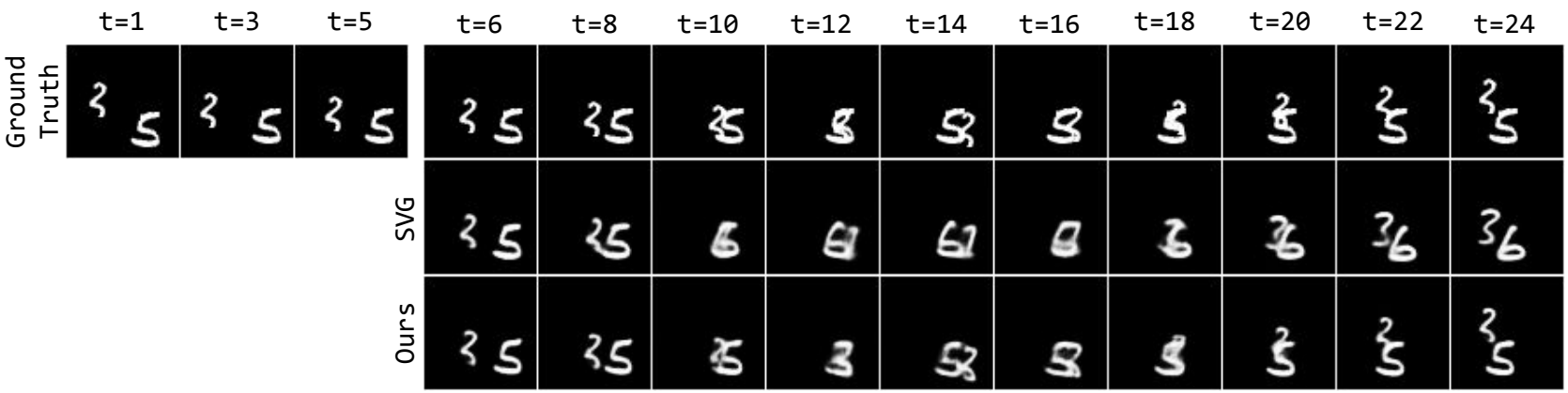
Stochastic Moving MNIST



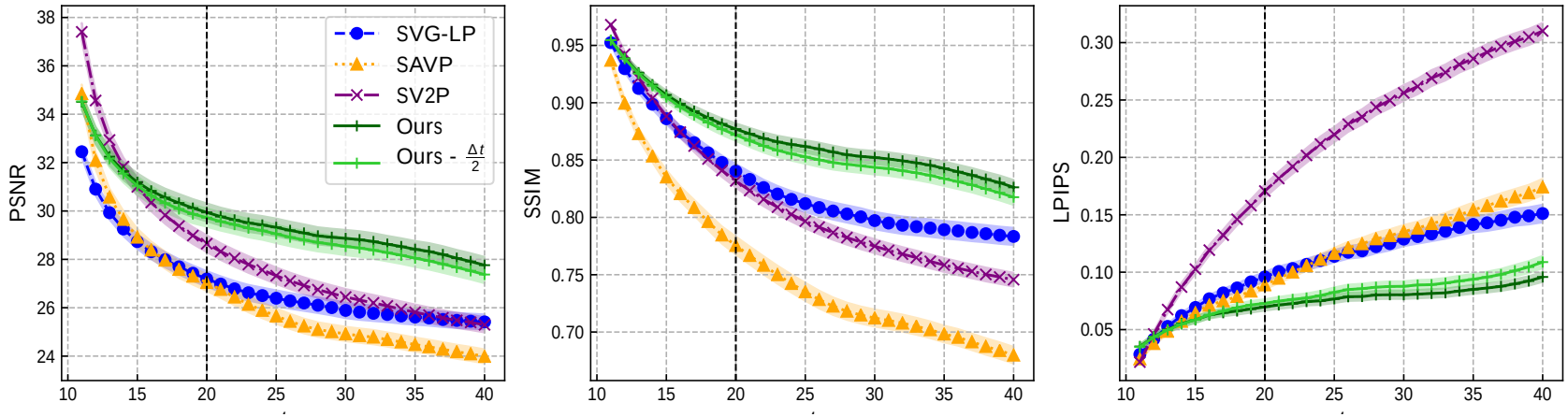
(a) Stochastic Moving MMNIST.

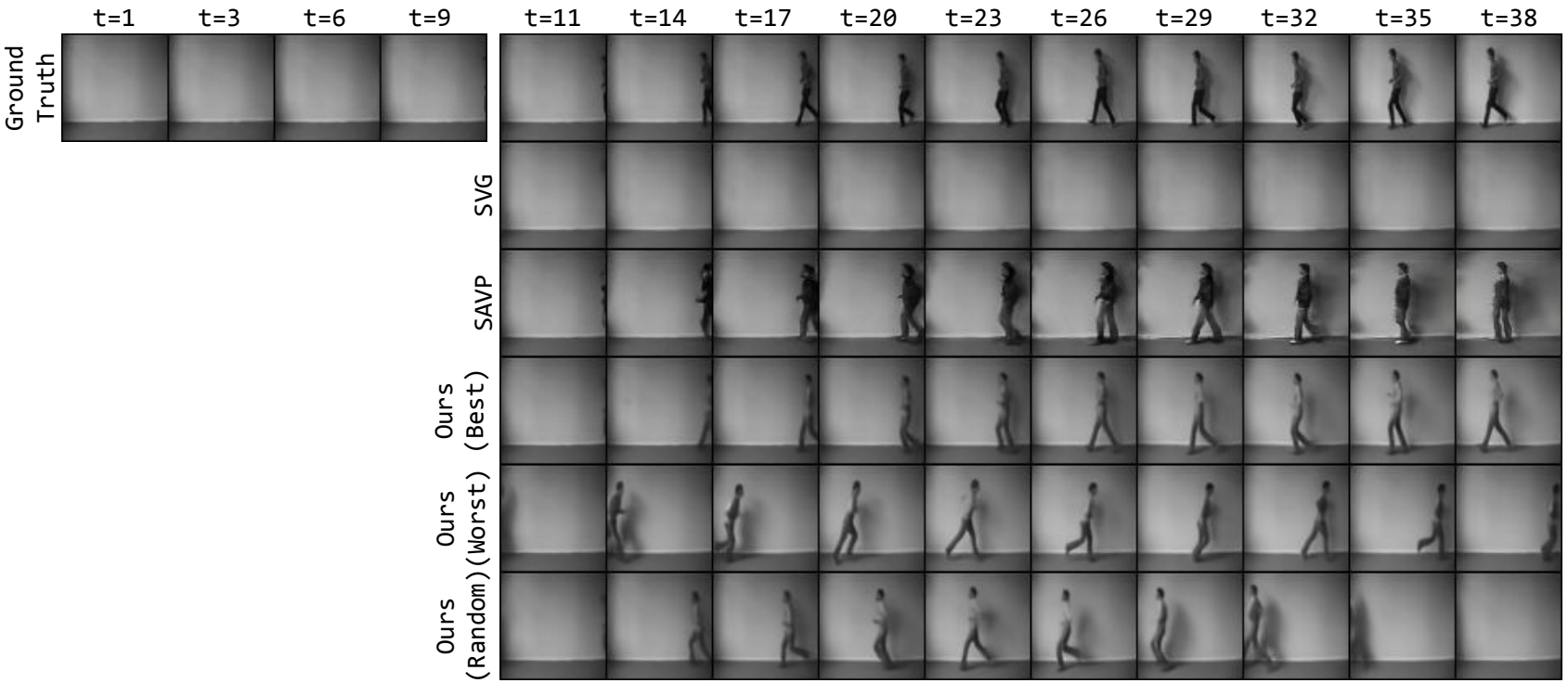


(b) Deterministic Moving MMNIST.

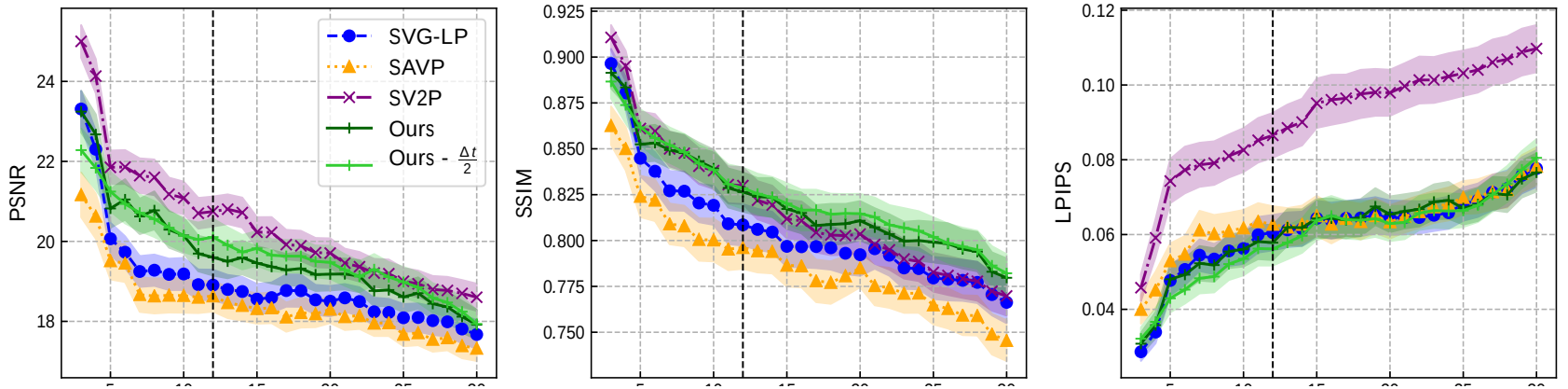


KTH Action Dataset





BAIR Robot Pushing Dataset



References

References I

- Babaeizadeh, Mohammad et al. (2018). “Stochastic Variational Video Prediction”. In: *International Conference on Learning Representations*.
- Chen, Tian Qi et al. (2018). “Neural Ordinary Di□erential Equations”. In: *Advances in Neural Information Processing Systems 31*. Ed. by Samy Bengio et al. Curran Associates, Inc., pp. 6571–6583.
- 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.
- Fraccaro, Marco et al. (2016). “Sequential Neural Models with Stochastic Layers”. In: *Advances in Neural Information Processing Systems 29*. Ed. by Daniel D. Lee et al. Curran Associates, Inc., pp. 2199–2207.
- 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.

References II

- Gregor, Karol et al. (2019). “Temporal Di□erence Variational Auto-Encoder”. In: *International Conference on Learning Representations*.
- Karl, Maximilian et al. (2017). “Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data”. In: *International Conference on Learning Representations)*.
- Krishnan, Rahul G., Uri Shalit, and David Sontag (2017). “Structured Inference Networks for Nonlinear State Space Models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31, pp. 2101–2109.
- Lee, Alex X. et al. (2018). “Stochastic Adversarial Video Prediction”. In: *arXiv preprint arXiv:1804.01523*.

Appendix

Prior

$$
\begin{cases}
y_1 \sim N(0; I) & \text{(initial condition)} \\
z_{t+1} \sim N(\mu(y_t); \sigma(y_t) I) & \text{(random variable prediction)} \\
y_{t+1} = y_t + f(y_t; z_{t+1}) & \text{(latent state prediction)} \\
x_t \sim N(g(y_t); I) & \text{(decoding)}
\end{cases}
$$

The above model induces the following factorization:

$$
\begin{aligned}
& p(x_{1:T}; z_{2:T}; y_{1:T} \mid w) \\
& \quad = p(y_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid y_t) p(y_{t+1} \mid y_t; z_{t+1}) \prod_{t=1}^{T} p(x_t \mid y_t; w);
\end{aligned}
$$

with $p(y_{t+1} \mid y_t; z_{t+1})$ being a Dirac.

Posterior

We factorize the posterior as follows:

$$q_{Z;Y} \;,\; q(z_{2:T}; y_{1:T} \; j \; x_{1:T}; w)$$

$$= q(y_1 \; j \; x_{1:k}) \prod_{t=2}^{T} q(z_t \; j \; x_{1:t}) \quad y_{t \;\; 1} + f \; (y_{t \;\; 1}; z_t) \quad y_t \; :$$

- y_1 is inferred from the first k frames
- z_t is inferred from the past and present frames only (works best in practice)
- y_t is deterministically computed from the previously inferred $y_{t \;\; 1}$ and z_t

Evidence Lower Bound

$$\log p(x_{1:T} \mid w) \quad L(x_{1:T}; w; \; ; \;)$$

$$, \; \mathrm{E}_{(\tilde{z}_{2:T}; \tilde{y}_{1:T}) \; q_{Z;Y}} \sum_{t=1}^{T} \log p(x_t \mid \hat{y}_t; w)$$

$$D_{\mathrm{KL}} \; q(y_1 \mid x_{1:k}) \quad p(y_1)$$

$$\mathrm{E}_{(\tilde{z}_{2:T}; \tilde{y}_{1:T}) \; q_{Z;Y}} \sum_{t=2}^{T} D_{\mathrm{KL}} \; q(z_t \mid x_{1:t}) \quad p(z_t \mid \hat{y}_{t \; 1})$$

Loss Function

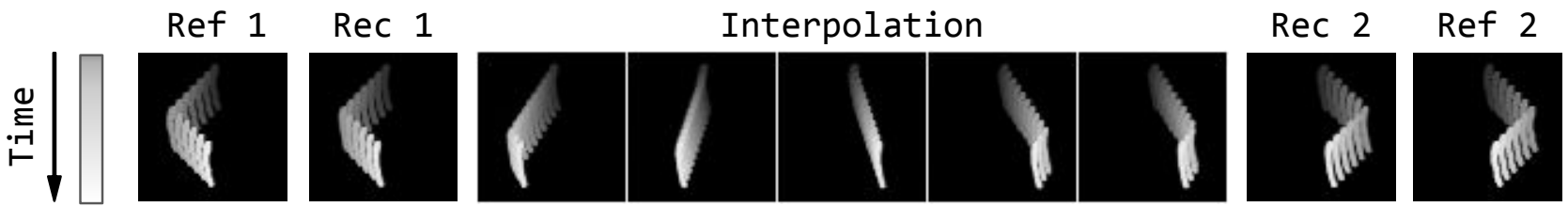
We add to the evidence lower bound an ℓ_2 regularization on the residues to stabilize the dynamics, giving the following loss function:

$$\arg\max_{\phi,\psi,\theta} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\boldsymbol{x}_c^{(k)}} \Big[L\left(\boldsymbol{x}_{1:T}, c\left(\boldsymbol{x}_c^{(k)}\right), \phi, \psi, \theta\right) - \mathbb{E}_{(\boldsymbol{z}_{2:T}, \boldsymbol{y}_{1:T}) \sim q_{Z,Y}} \Big[\sum_{t=2}^{T} \left\| f_\theta(y_{t-1}, z_t) \right\|_2^2 \Big] \Big].$$

Latent Space Interpolation

Interpolation of y_1 on Moving MNIST

- Take 2 trajectories x^s and x^t
- Infer Latent initial conditions y_1^s and y_1^t
- Generate frame sequences from $\tilde{y}_1$ linearly interpolated between y_1^s and y_1^t.



Pendulum

Deep State Space Model

Without content variable, we can compare our model to Deep State Space Models.

Table 1: ELBO for DVBF [Kar+17], KVAE [Fra+17] and our model on the Pendulum dataset.

Score	DVBF	KVAE	Ours
ELBO	798.56	807.02	806.12