Spatio-temporal and Video Prediction with State-based Neural Models

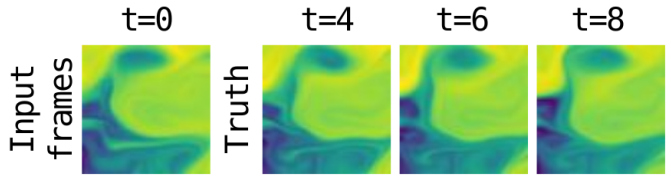
February 1st, 2022

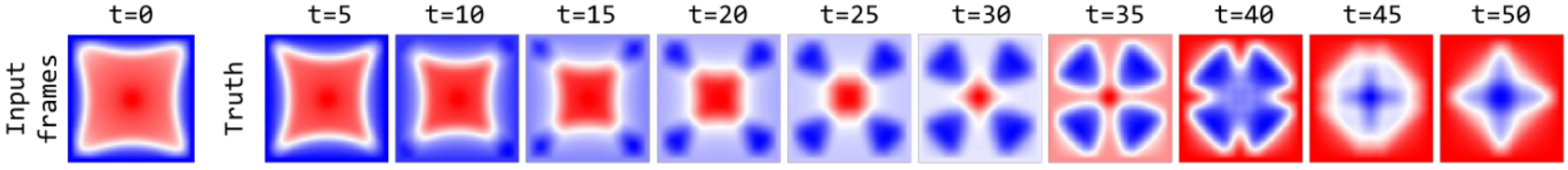
Jean-Yves Franceschi,[1] Edouard Delasalles,[1]
Jérémie Donà,[1] Mickael Chen,[1]
Sylvain Lamprier,[1] Patrick Gallinari[1,2]

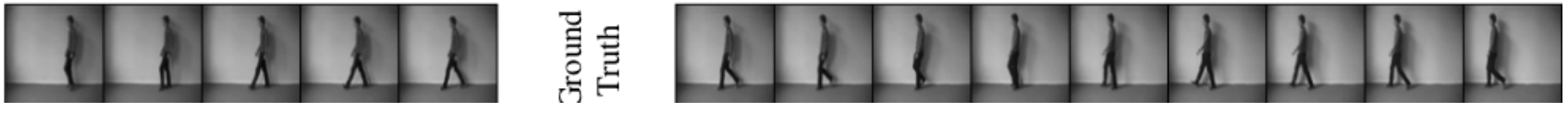
[1]Sorbonne Université, CNRS, ISIR, Paris, France
[2]Criteo AI Lab, Paris, France

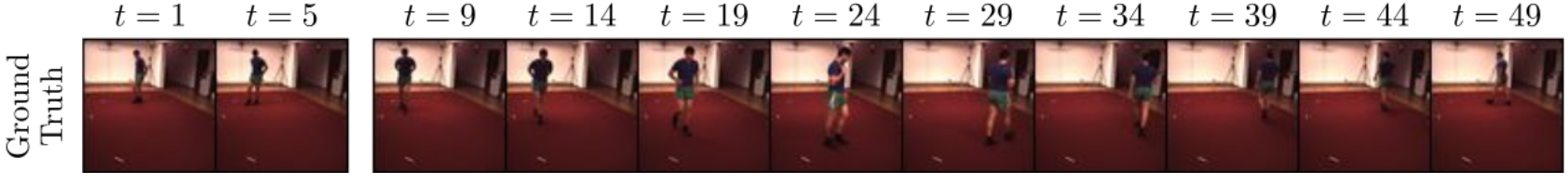
Machine Learning & Deep Learning for Information Access

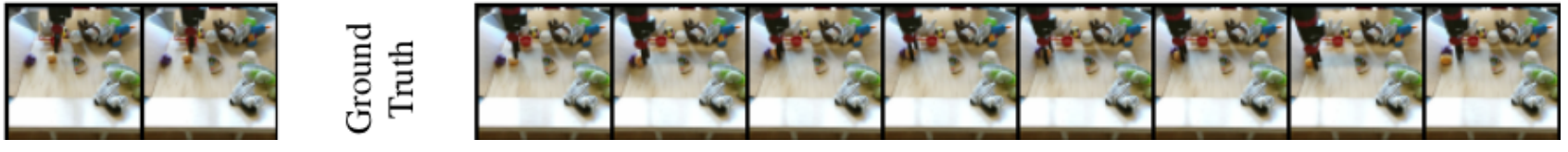
Given conditioning frames, predict the distribution of future frames.











Given conditioning frames, predict the distribution of future frames.

Applications

- Visual Forecast, Analysis (e.g., climate, health, crowds, etc.)
- Model-based Reinforcement Learning (Gregor et al., 2019)

Challenges

- Capture visual and dynamic representations of the world
- Generation of realistic images
- Long-term prediction (not only at t+1)
- Account for uncertainty in the future
- Disentangling static/dynamic objects

Given conditioning frames, predict the distribution of future frames.

Applications

- Visual Forecast, Analysis (e.g., climate, health, crowds, etc.)
- Model-based Reinforcement Learning (Gregor et al., 2019)

Challenges

- Capture visual and dynamic representations of the world
- Generation of realistic images
- Long-term prediction (not only at t+1)
- **Account for uncertainty in the future**
- Disentangling static/dynamic objects

Given conditioning frames, predict the distribution of future frames.

Applications

- Visual Forecast, Analysis (e.g., climate, health, crowds, etc.)
- Model-based Reinforcement Learning (Gregor et al., 2019)

Challenges

- Capture visual and dynamic representations of the world
- Generation of realistic images
- Long-term prediction (not only at t+1)
- Account for uncertainty in the future
- **Disentangling static/dynamic objects**

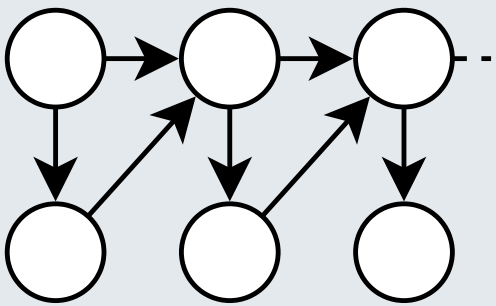
Stochastic Video Prediction

Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, Patrick Gallinari: Stochastic Latent Residual Video Prediction. ICML 2020: 3233-3246

Autoregressive Models

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

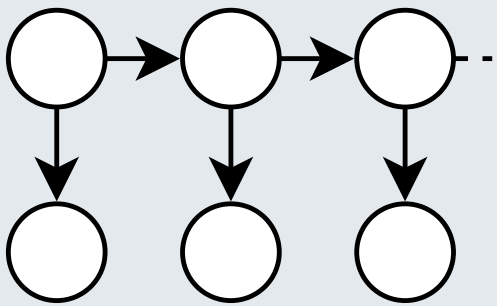
Ex.: Denton and Fergus (2018)



State-Space Models

+ Decoupled dynamics and prediction, interpretable
– Open-loop: Harder to train

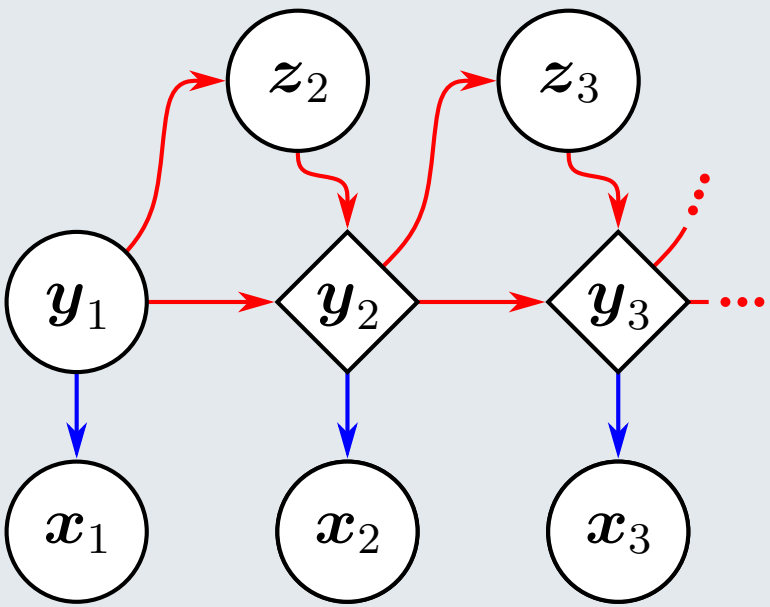
Ex.: Fraccaro et al. (2017)



State-space Model with Residual Updates

- High-order Dynamics in state y_t (memory)
- From 1st order ODE inspiration : $\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = f_{\boldsymbol{z}_{\lfloor t \rfloor + 1}}(\boldsymbol{y})$
- State-dependent noise z_t (stochasticity not only on y_1)

Model

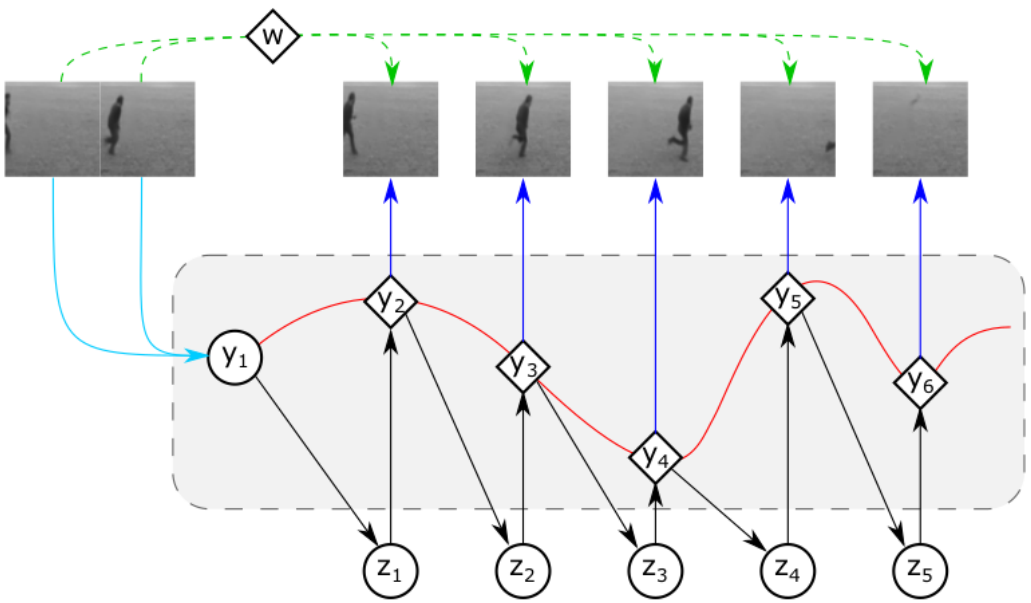


Update Rule:

$$\begin{cases} \boldsymbol{z}_{t+1} \sim \mathcal{N}\big(\mu_\theta(\boldsymbol{y}_t), \sigma_\theta(\boldsymbol{y}_t) I\big) \\ \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + f_\theta(\boldsymbol{y}_t, \boldsymbol{z}_{t+1}) \end{cases}$$

State-space Model with Residual Updates

- Variable w to Separate stochasticity from dynamics
  - Better gradient backprop



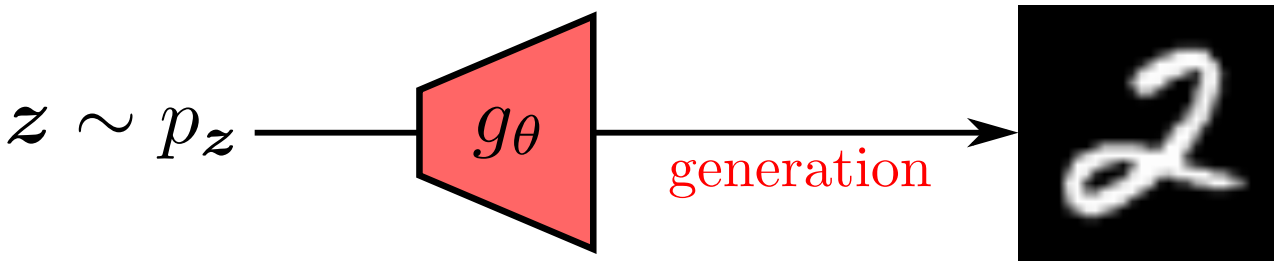
$$\mathbf{y}_{t+1} = \mathbf{y}_t + f_\theta(\mathbf{y}_t, \mathbf{z}_{t+1})$$

$$\mathbf{z}_{t+1} \sim \mathcal{N}\left(\mu_\theta(\mathbf{y}_t), \sigma_\theta(\mathbf{y}_t) I\right)$$

- Generative model $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ for $\boldsymbol{x}$ from a latent $\boldsymbol{z} \sim p_\theta(\boldsymbol{z})$:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$$

- Intractable objective: sampling from $p(z)$ very inefficient



- Introducing $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to approximate $p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) \propto p_\theta(\boldsymbol{z}, \boldsymbol{x})$:

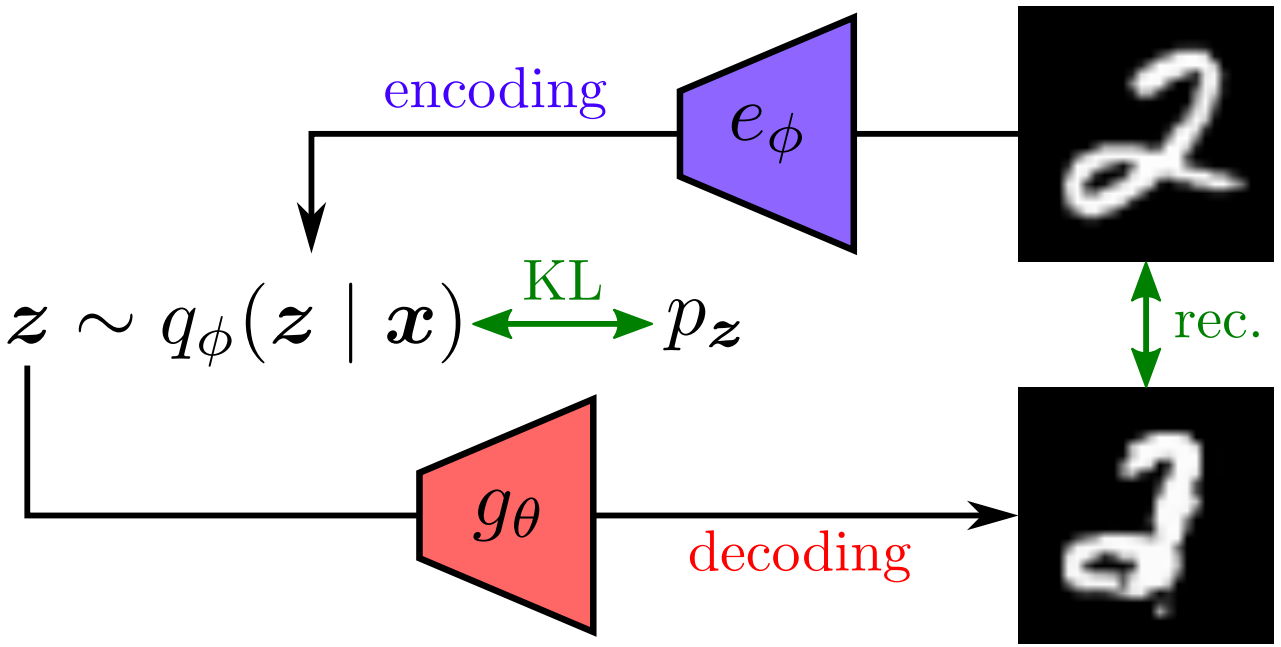
$$D_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\big\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})\big) = p_\theta(\boldsymbol{x}) - \int q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{z}, \boldsymbol{x})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} dz$$

$$\Rightarrow \arg\min_\phi D_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\big\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})\big) = \arg\max_\phi \int q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{z}, \boldsymbol{x})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} dz$$

Evidence Lower Bound (ELBO)

Optimized objective:

$$\arg\max_{\phi,\theta} \underbrace{\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\middle\|\,p(\boldsymbol{z})\right)}_{\text{KL term}}.$$

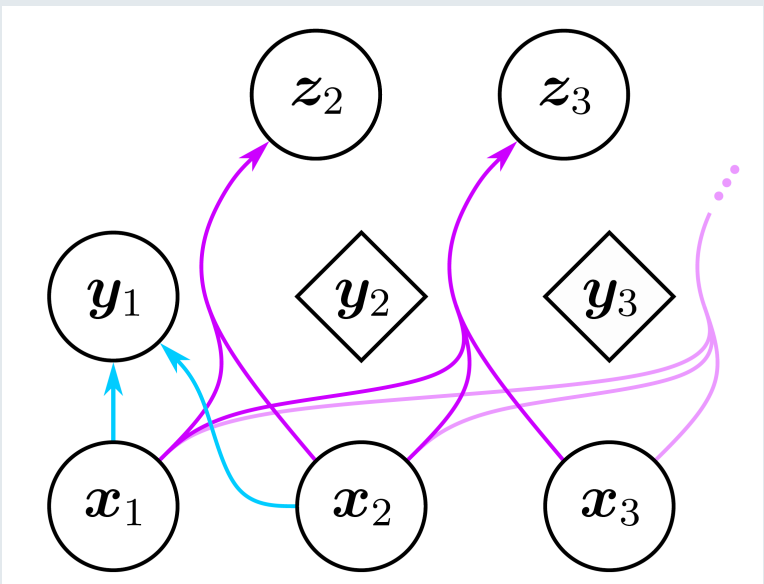


State-space Model with Residual Updates

- Initial Condition y_1 unknown, Stochastic Noise z unknown

$\Rightarrow$ $p(\boldsymbol{y}_1, \boldsymbol{z}_{2:T} \mid \boldsymbol{x}_{1:T})$ to be infered via VAE

Inference

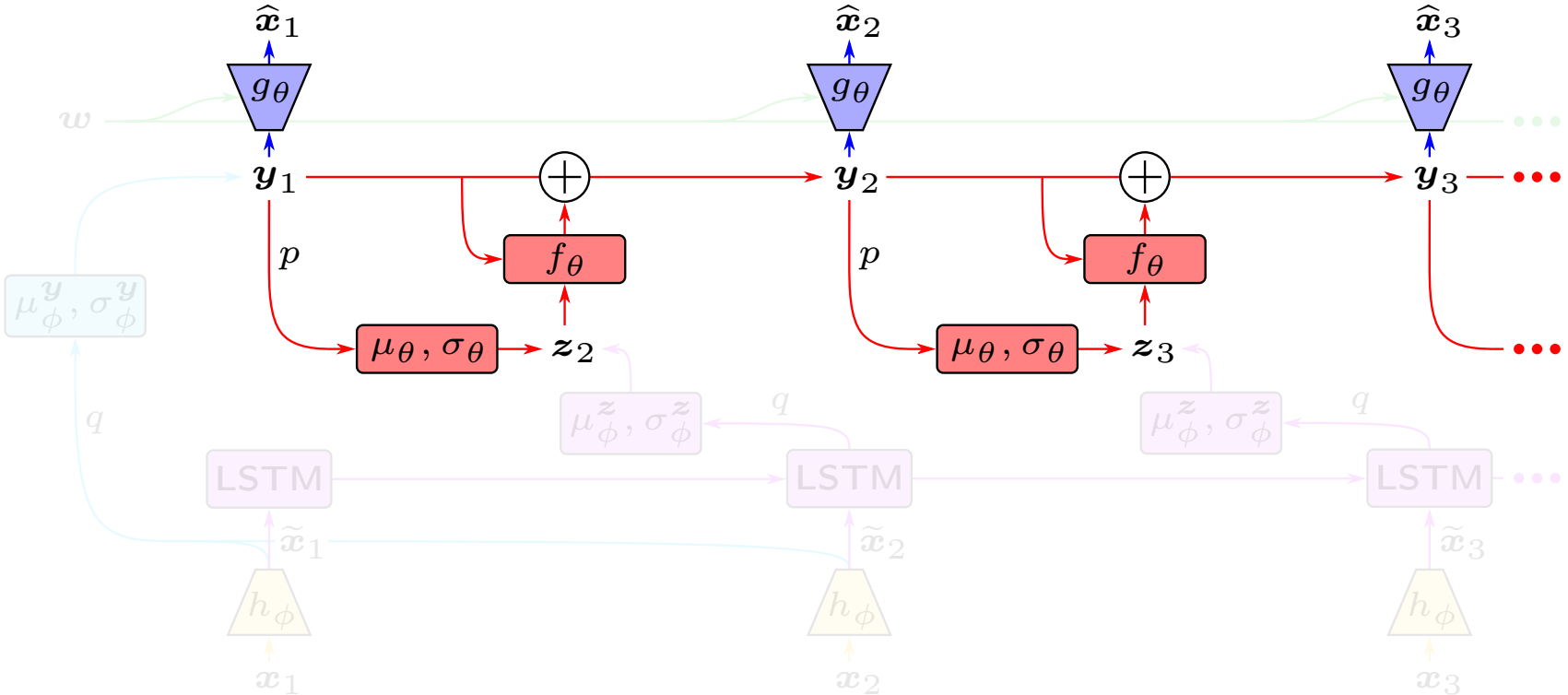


Variational Distribution:

$$q(\boldsymbol{y}_1, \boldsymbol{z}_{2:T} \mid \boldsymbol{x}_{1:T}) = q_\phi^y(\boldsymbol{y}_1 \mid \boldsymbol{x}_{1:k}) \prod_{t=2}^{T} q_\phi^z(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})$$

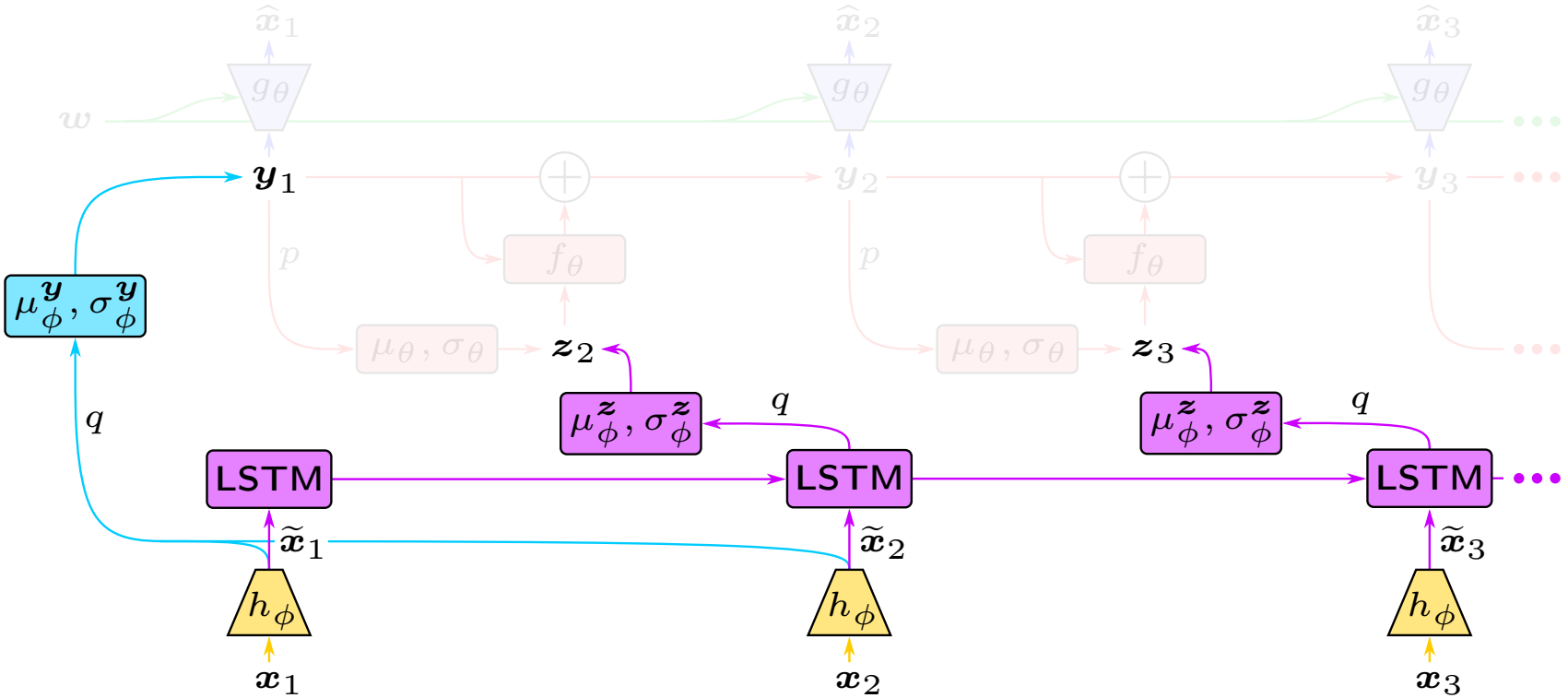
$$\begin{cases} \boldsymbol{y}_1 \sim \mathcal{N}\left(\mu_\phi^y(\boldsymbol{x}_{1:k}), \sigma_\phi^y(\boldsymbol{x}_{1:k}) I\right) \\ \boldsymbol{z}_t \sim \mathcal{N}\left(\mu_\phi^z(\boldsymbol{x}_{1:t}), \sigma_\phi^z(\boldsymbol{x}_{1:t}) I\right) \end{cases}$$

Generative Model

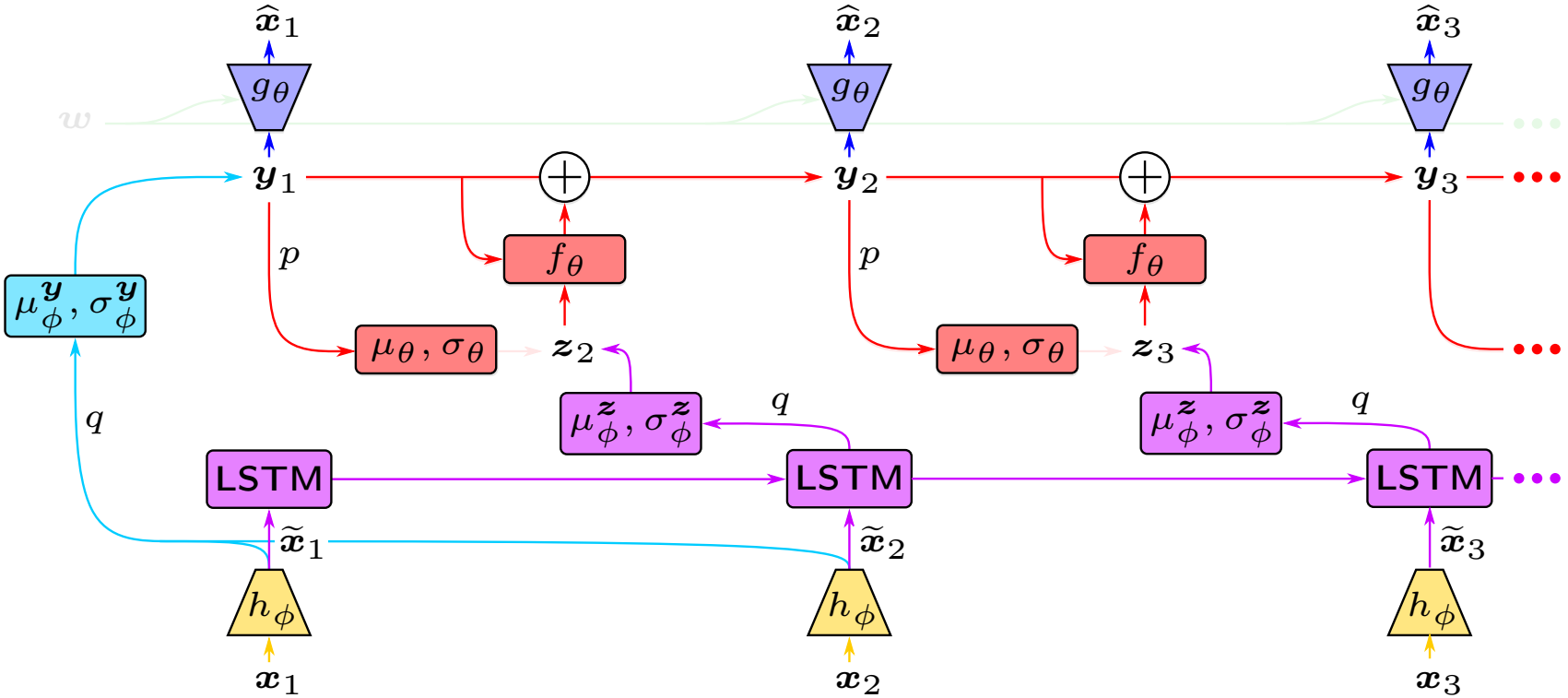


$$
\begin{cases}
\boldsymbol{y}_1 \sim \mathcal{N}(\mathbf{0}, I) & \text{(initial condition)} \\
\boldsymbol{z}_{t+1} \sim \mathcal{N}\big(\mu_\theta(\boldsymbol{y}_t), \sigma_\theta(\boldsymbol{y}_t) I\big) & \text{(random variable prediction)} \\
\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + f_\theta(\boldsymbol{y}_t,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

Inference



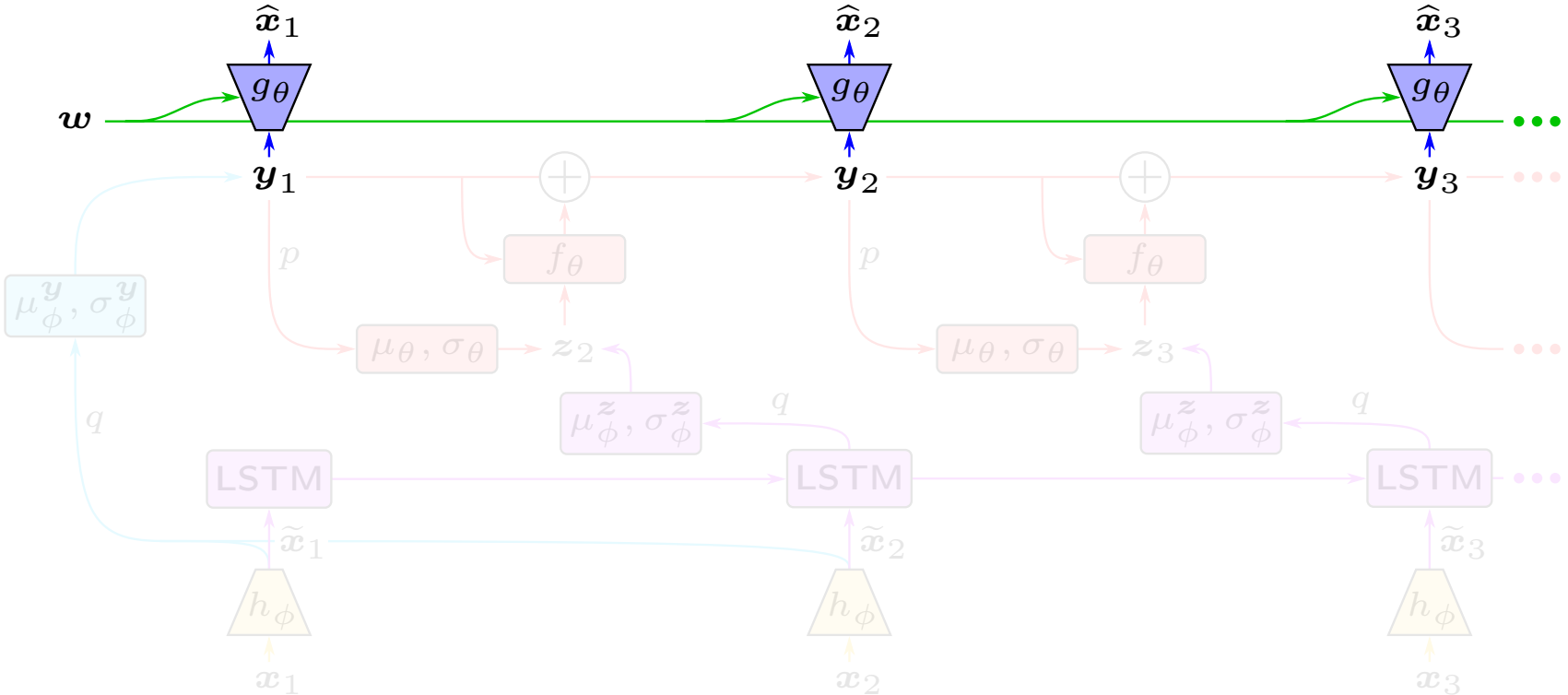
$$q(\boldsymbol{z}_{2:T}, \boldsymbol{y}_1 \mid \boldsymbol{x}_{1:T}) = \underbrace{q(\boldsymbol{y}_1 \mid \boldsymbol{x}_{1:k})}_{\text{Init. Cond.}} \prod_{t=2}^{T} \underbrace{q(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})}_{\text{LSTM}}$$



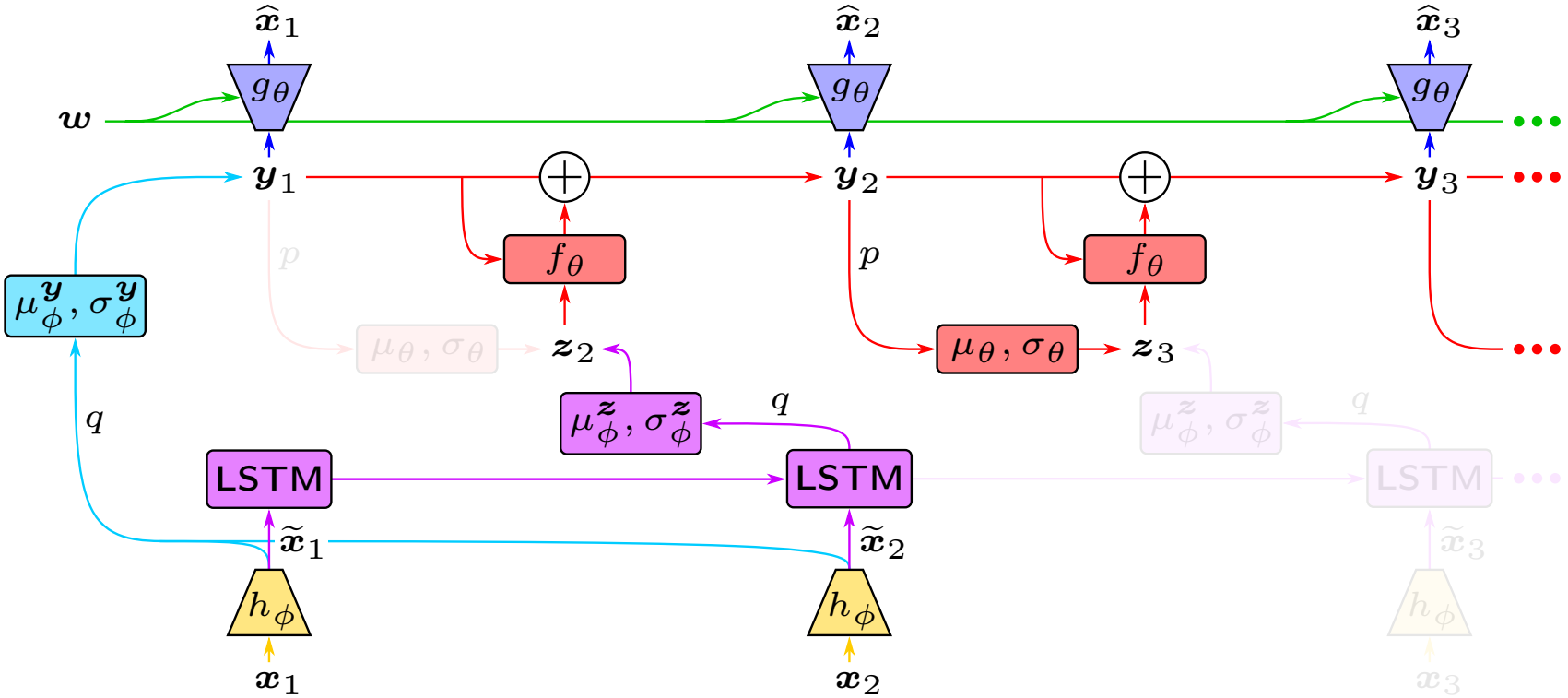
$$q(\boldsymbol{z}_{2:T}, \boldsymbol{y}_1 \mid \boldsymbol{x}_{1:T}) = \underbrace{q(\boldsymbol{y}_1 \mid \boldsymbol{x}_{1:k})}_{\text{Init. Cond.}} \prod_{t=2}^{T} \underbrace{q(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})}_{\text{LSTM}}$$

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

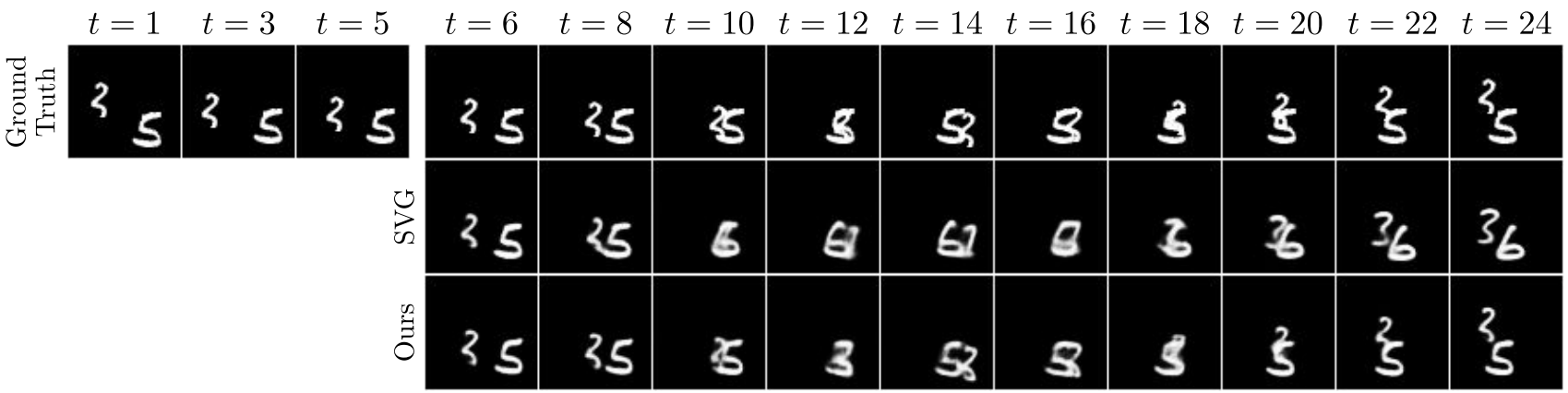


- Conditioning frames are used to infer dynamic variables
- Prediction follows using the forward model

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)

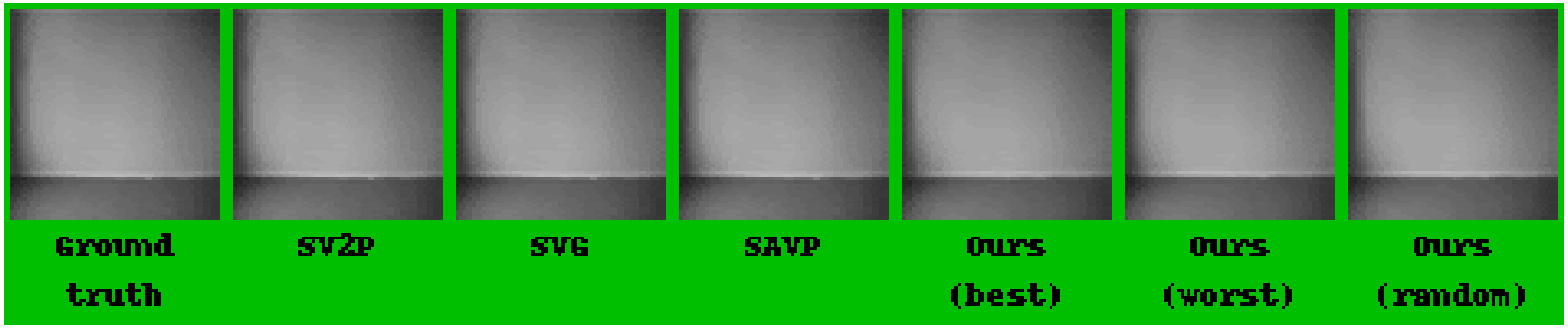


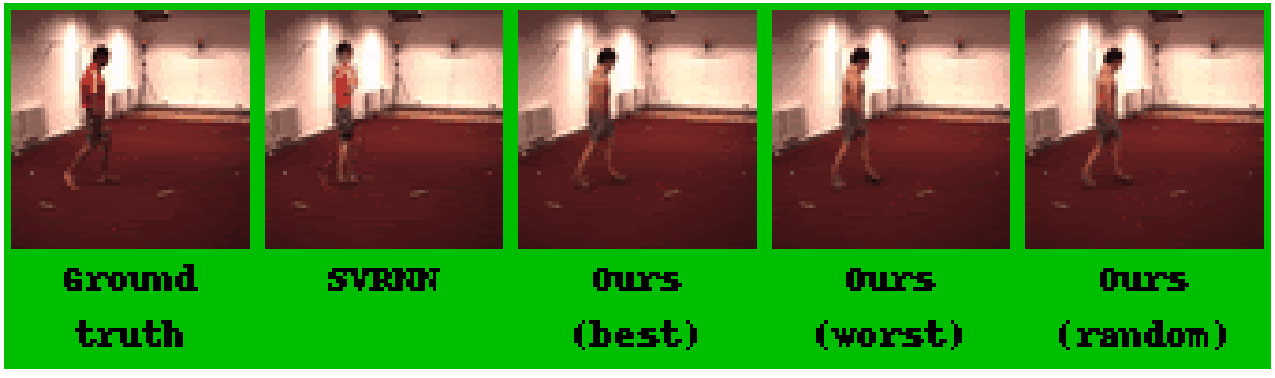
Experimental Results: KTH, Human3.6M and BAIR

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 et al. (2016), SAVP: Babaeizadeh et al. (2018), SVRNN: Minderer et al. (2019)

Experimental Results: Samples

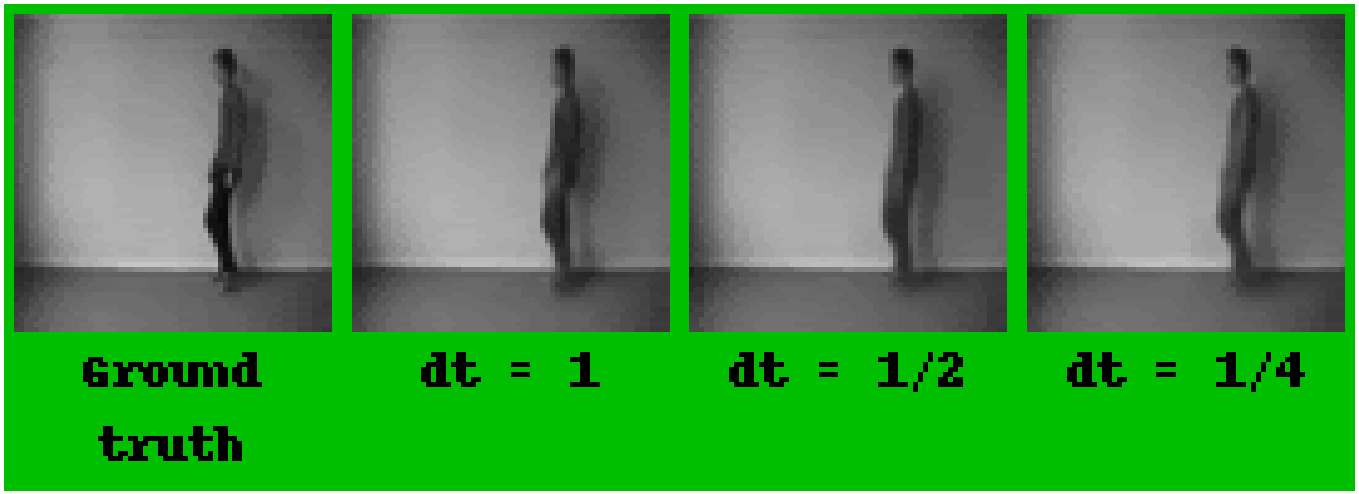


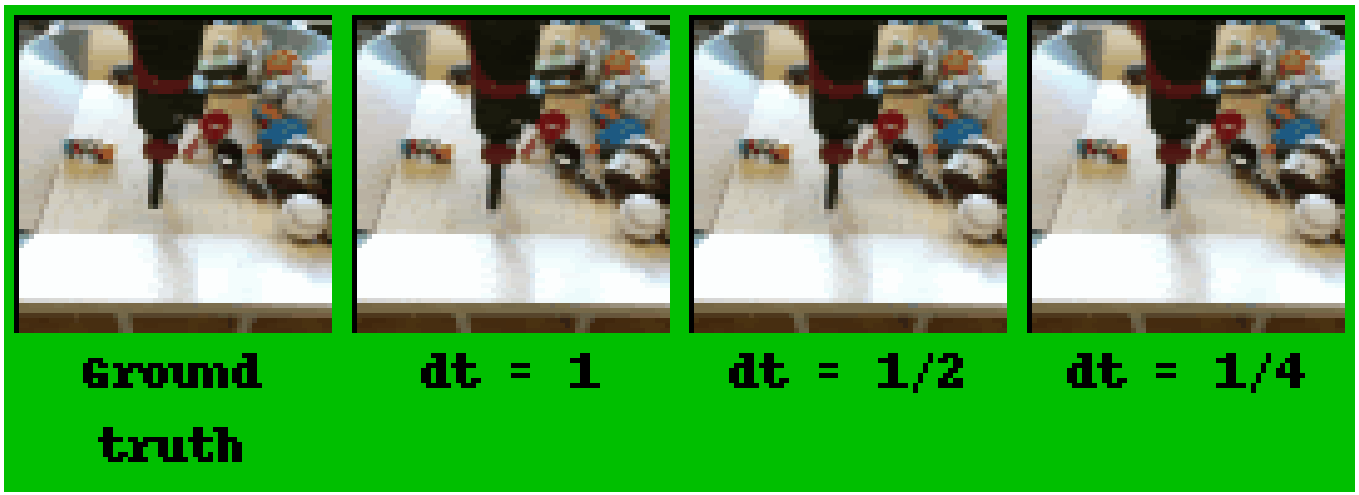


Experimental Results: Generation at Various Frame Rates

Using the Euler approximation scheme:

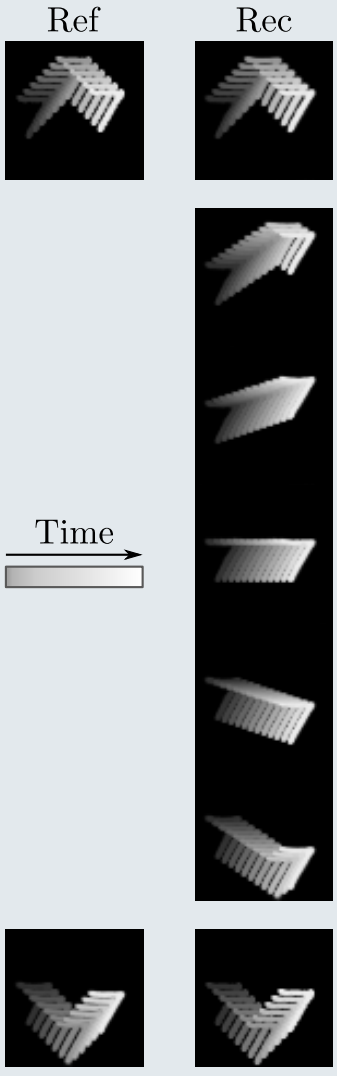
$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = f_{\boldsymbol{z}_{\lfloor t \rfloor + 1}}(\boldsymbol{y}) \quad \Rightarrow \quad \boldsymbol{y}_{t+\Delta t} = \boldsymbol{y}_t + \Delta t \cdot f_\theta\Big(\boldsymbol{y}_t, \boldsymbol{z}_{\lfloor t \rfloor + 1}\Big)$$



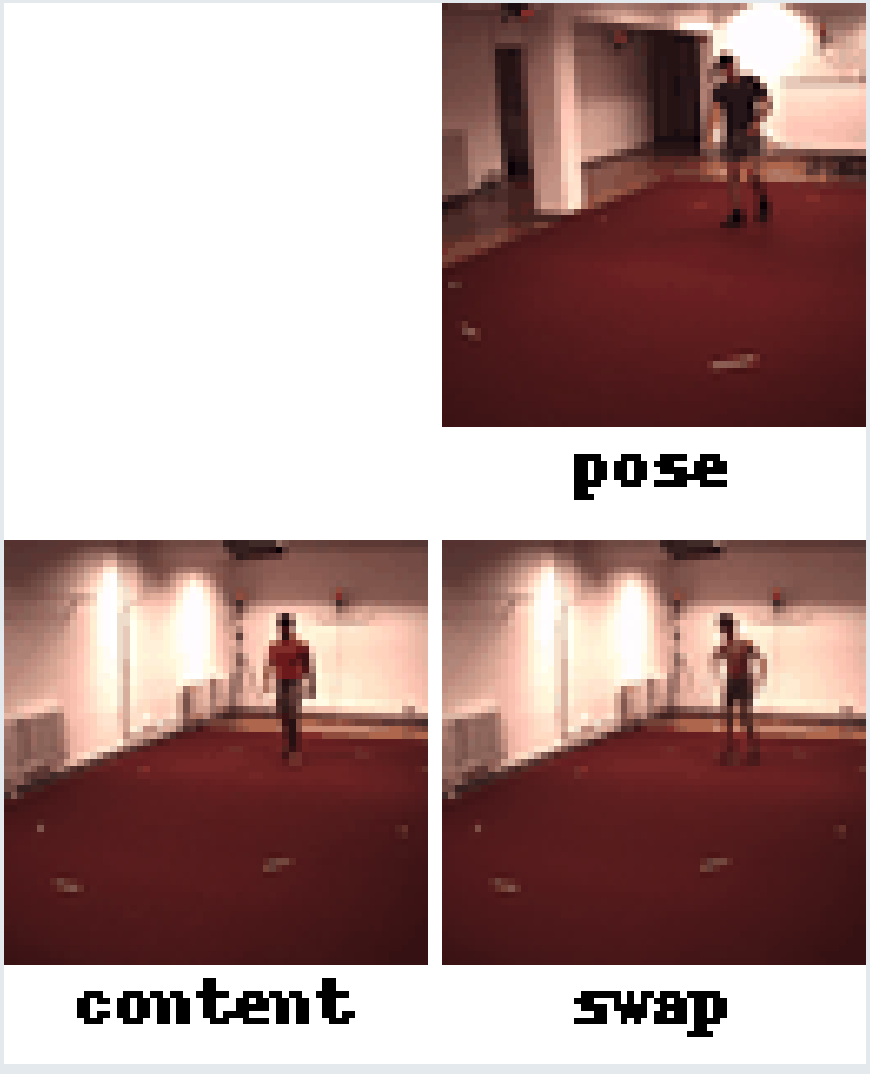


Experimental Results: Latent Space Properties

Dynamics interpolation



Content Swap



- The separation of dynamic variables $\boldsymbol{y}$ and the content variable $\boldsymbol{w}$ can be seen as spatiotemporal disentanglement.
- Spatiotemporal disentanglement provides interpretability.
- It can help to improve prediction performance.

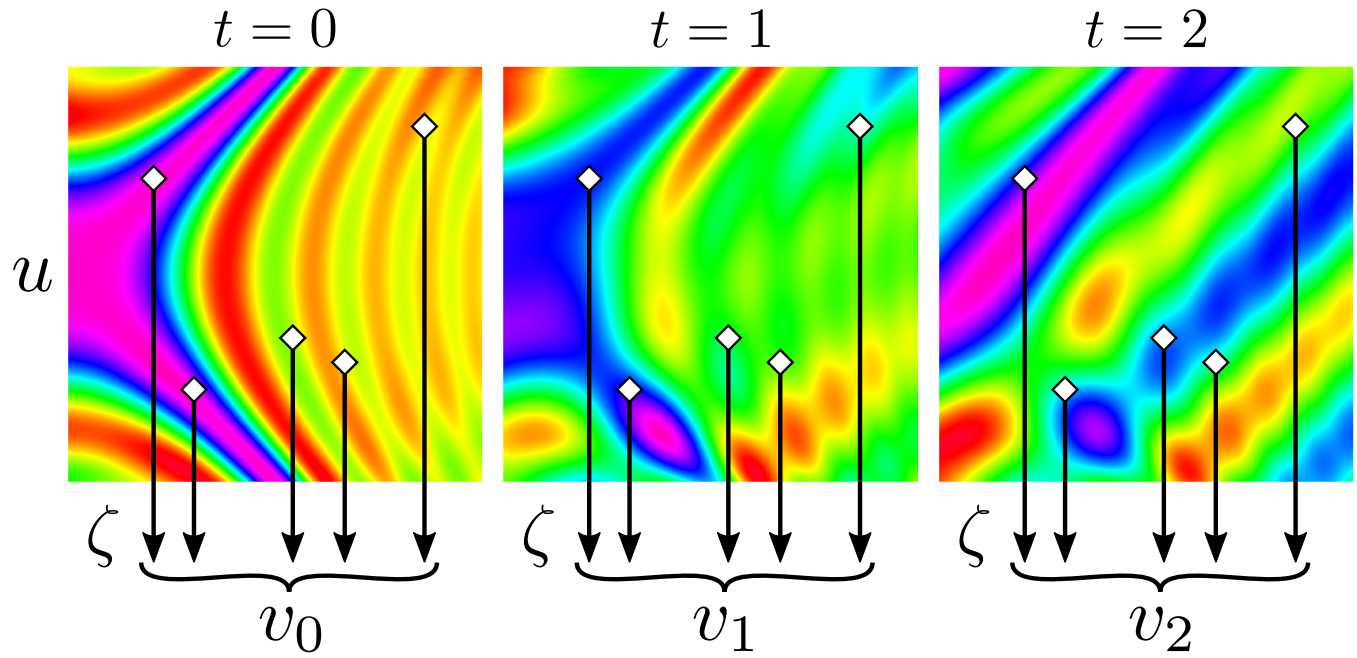
Question

What is spatiotemporal disentanglement?

- Often complex and seldom analyzed, achieved through:
  - KL-based separation in VAEs: ours, Hsieh et al. (2018) and Yingzhen et al. (2018);
  - adversarial losses: Villegas et al. (2017) and Denton and Birodkar (2017).
- We aim at grounding spatiotemporal disentanglement on stronger foundations, with fewer implicit assumptions

Spatiotemporal Disentanglement

Jérémie Donà, Jean-Yves Franceschi, Sylvain Lamprier, and Patrick Gallinari: PDE-Driven Spatiotemporal Disentanglement. ICLR 2021.



Hidden State u

Function of continuous coordinates following a PDE:

$$u\colon (x,t) \mapsto u(x,t).$$

Ex.: physical state of an action, ocean temperature.

Observations $\boldsymbol{v}$

Vectorial $\boldsymbol{v}_{t_0}, \ldots, \boldsymbol{v}_{t_1}$, spatial measurements of u:

$$\boldsymbol{v}_t = \zeta \circ u(\cdot, t).$$

Ex.: pixel values, punctual surface temperatures.

Idea: Separation of Variables

- We propose a novel interpretation of spatiotemporal disentanglement based on the separation of variables in PDEs
- Example with the heat equation:

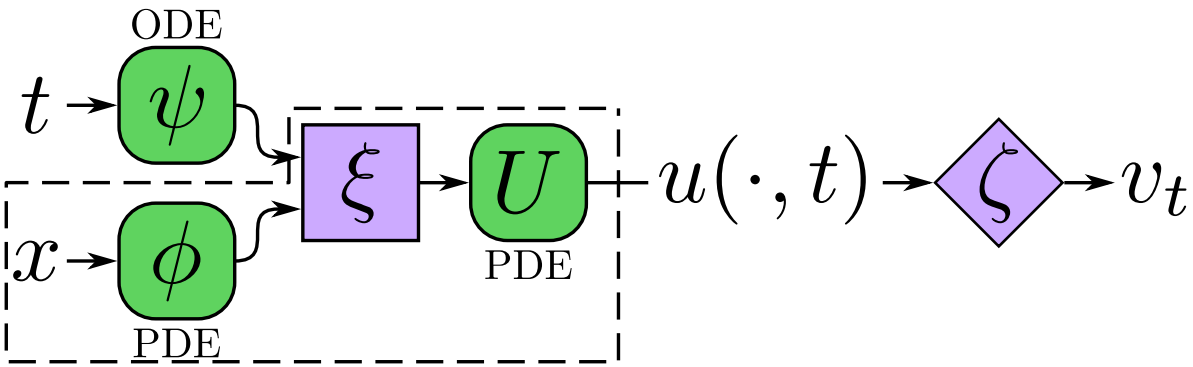
$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad u(0,t) = u(L,t) = 0, \quad u(x,0) = f(x),$$

with separable solutions:

$$u(x,t) = \underbrace{\mu \sin\left(\frac{n\pi}{L}x\right)}_{\phi(x)} \times \underbrace{\exp\left(-\left(\frac{cn\pi}{L}\right)^2 t\right)}_{\psi(t)} = \xi\big(\phi(x), \psi(t)\big)$$

- ϕ and ψ can be found by solving an ODE on t and a PDE on x

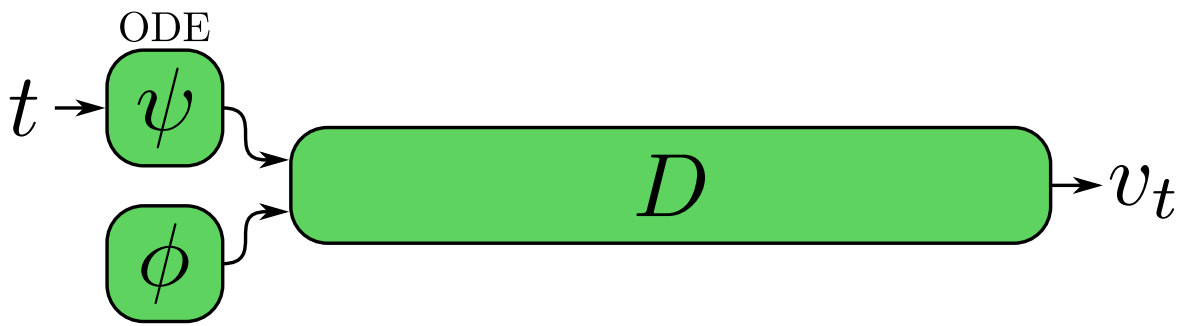
For spatiotemporal modeling from observations v:



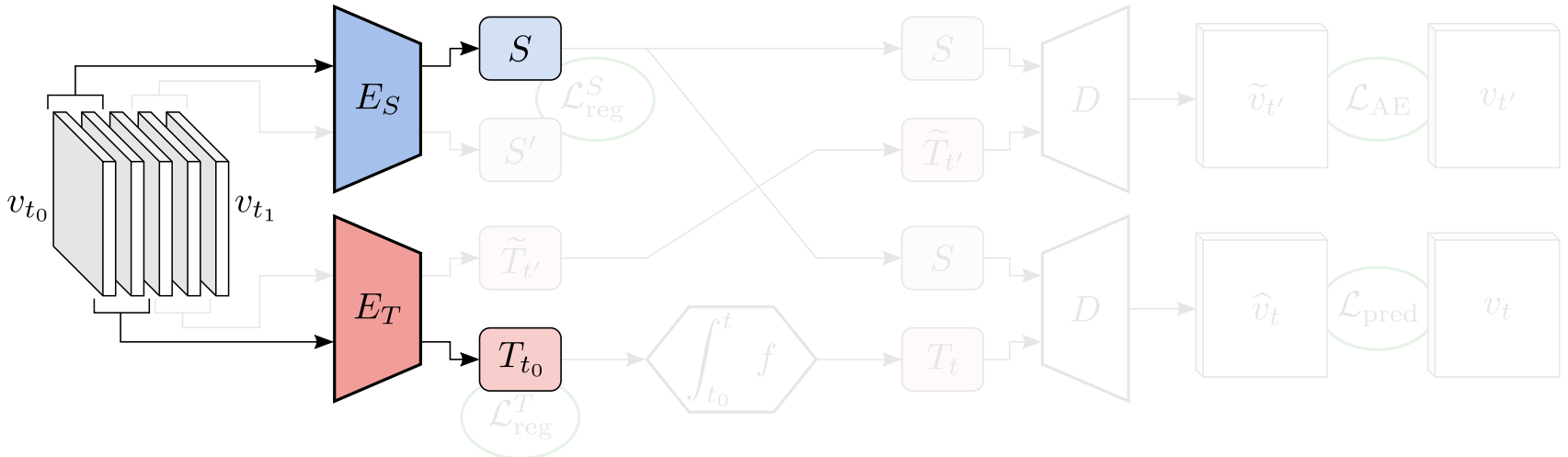
But directly learning U from partial observations can be too difficult:

- No always access to coordinates x
- Many unobserved variables

Thus, focus on learning ϕ, ψ and decoder $D\big(\phi, \psi(t)\big) = (\zeta \circ U \circ \xi)\big(\phi(\cdot), \psi(t)\big)$:



Model Variables

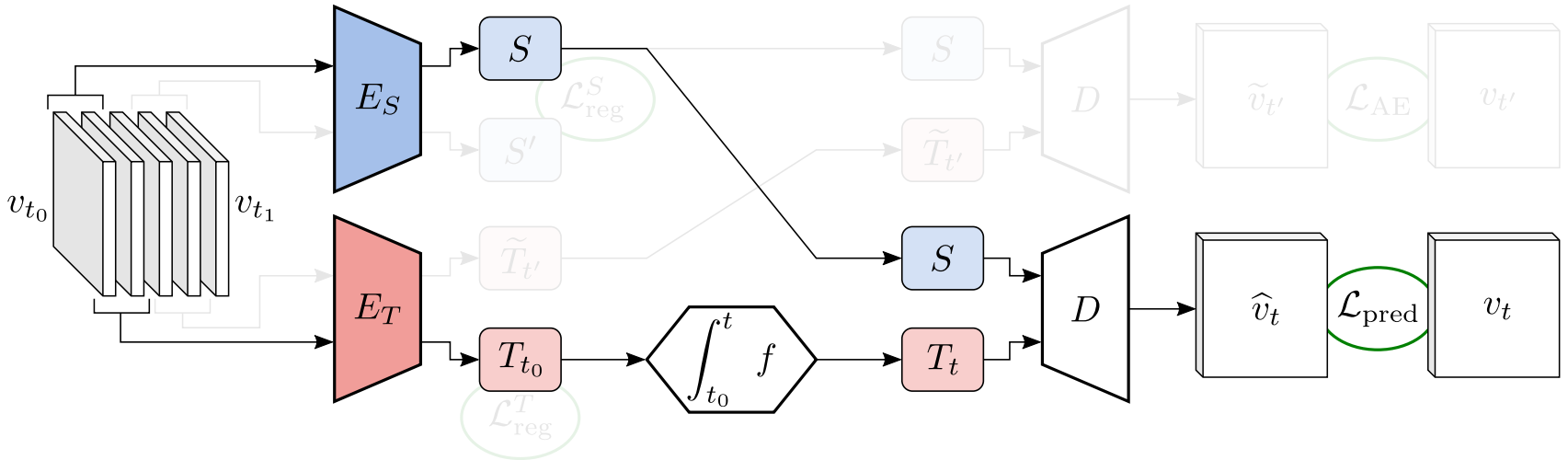


- We learn representations S, T_t such that:

$$\phi \equiv S \in \mathbb{R}^d, \qquad \psi \equiv T : t \mapsto T_t \in \mathbb{R}^p$$

- S and T_{t_0} are inferred with encoders E_S and E_T from conditioning frames:

$$V_\tau(t_0) = \left(v_{t_0}, \ldots, v_{t_0+\tau}\right)$$

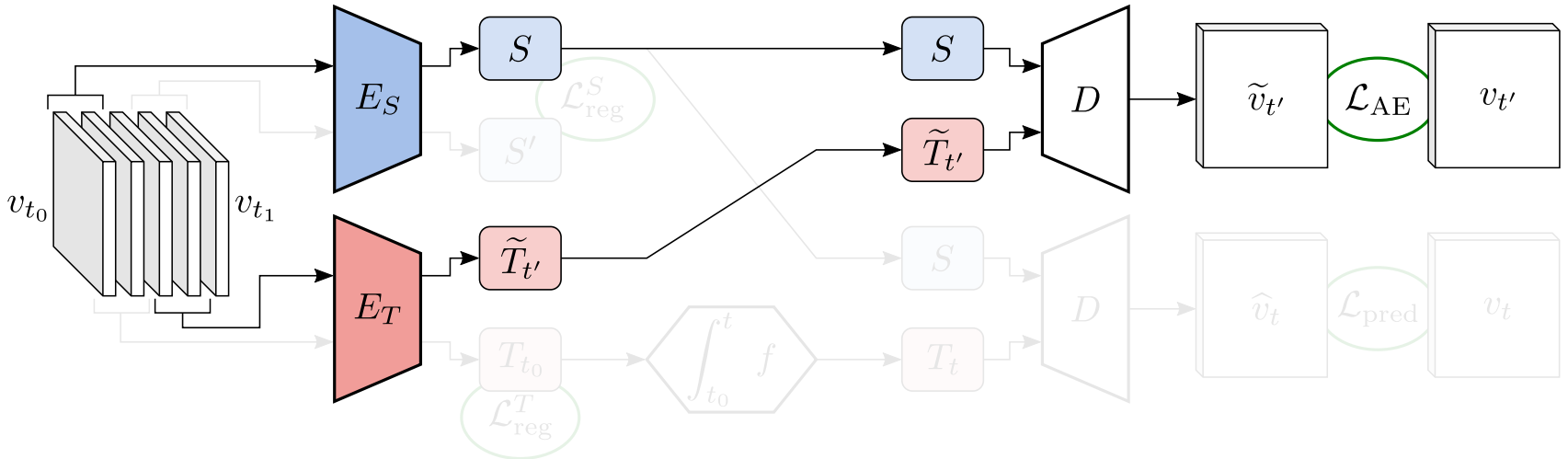


- $T \equiv \psi$ is driven by an ODE:

$$\frac{\partial T_t}{\partial t} = f(T_t) \qquad \Leftrightarrow \qquad T_t = T_{t_0} + \int_{t_0}^{t} f(T_{t'})\,\mathrm{d}t'$$

- Forecasting loss:

$$\widehat{v}_t = D\left(S, T_{t_0} + \int_{t_0}^{t} f(T_{t'})\,\mathrm{d}t'\right), \quad \mathcal{L}_{\text{pred}} = \frac{1}{\nu+1}\sum_{i=0}^{\nu}\frac{1}{m}\left\|\widehat{v}_{t_0+i\Delta t} - v_{t_0+i\Delta t}\right\|_2^2$$



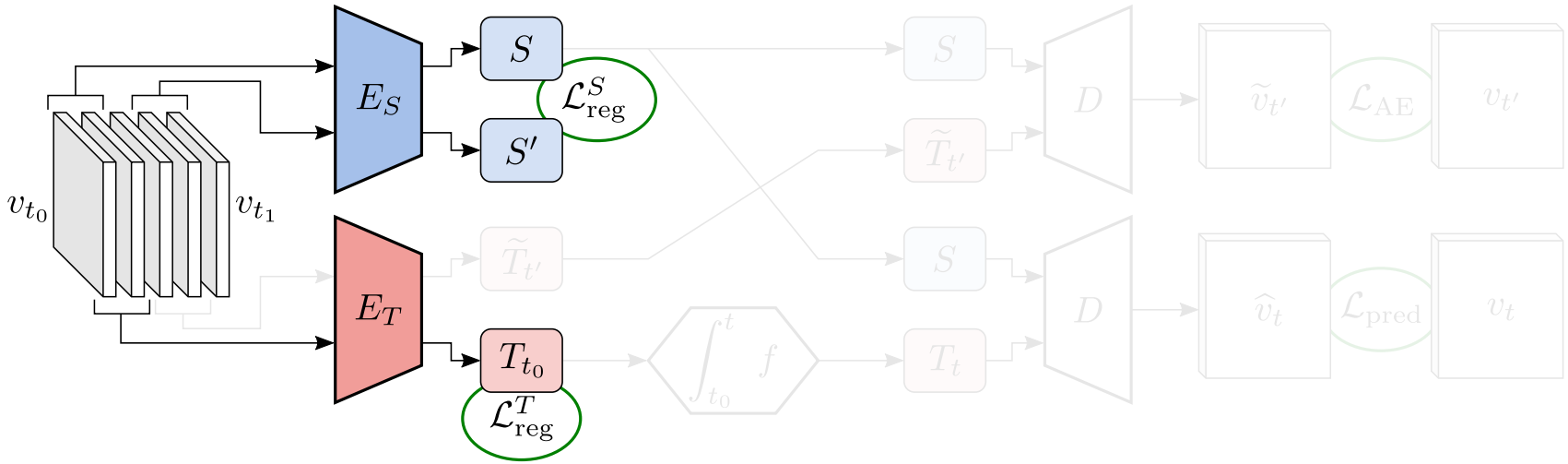
- $T \equiv \psi$ is driven by an ODE:

$$\frac{\partial T_t}{\partial t} = f(T_t) \qquad \Leftrightarrow \qquad T_t = T_{t_0} + \int_{t_0}^{t} f(T_{t'})\,\mathrm{d}t'$$

- Alignment loss:

$$\mathcal{L}_{\text{AE}} = \frac{1}{m}\left\| D\Big(S, E_T\big(V_\tau(t_0 + i\Delta t)\big)\Big) - v_{t_0+i\Delta t} \right\|_2^2, \quad \text{with } i \sim \mathcal{U}\big(\llbracket 0, \nu - \tau \rrbracket\big).$$

Invariance of S and Spatiotemporal Disentanglement



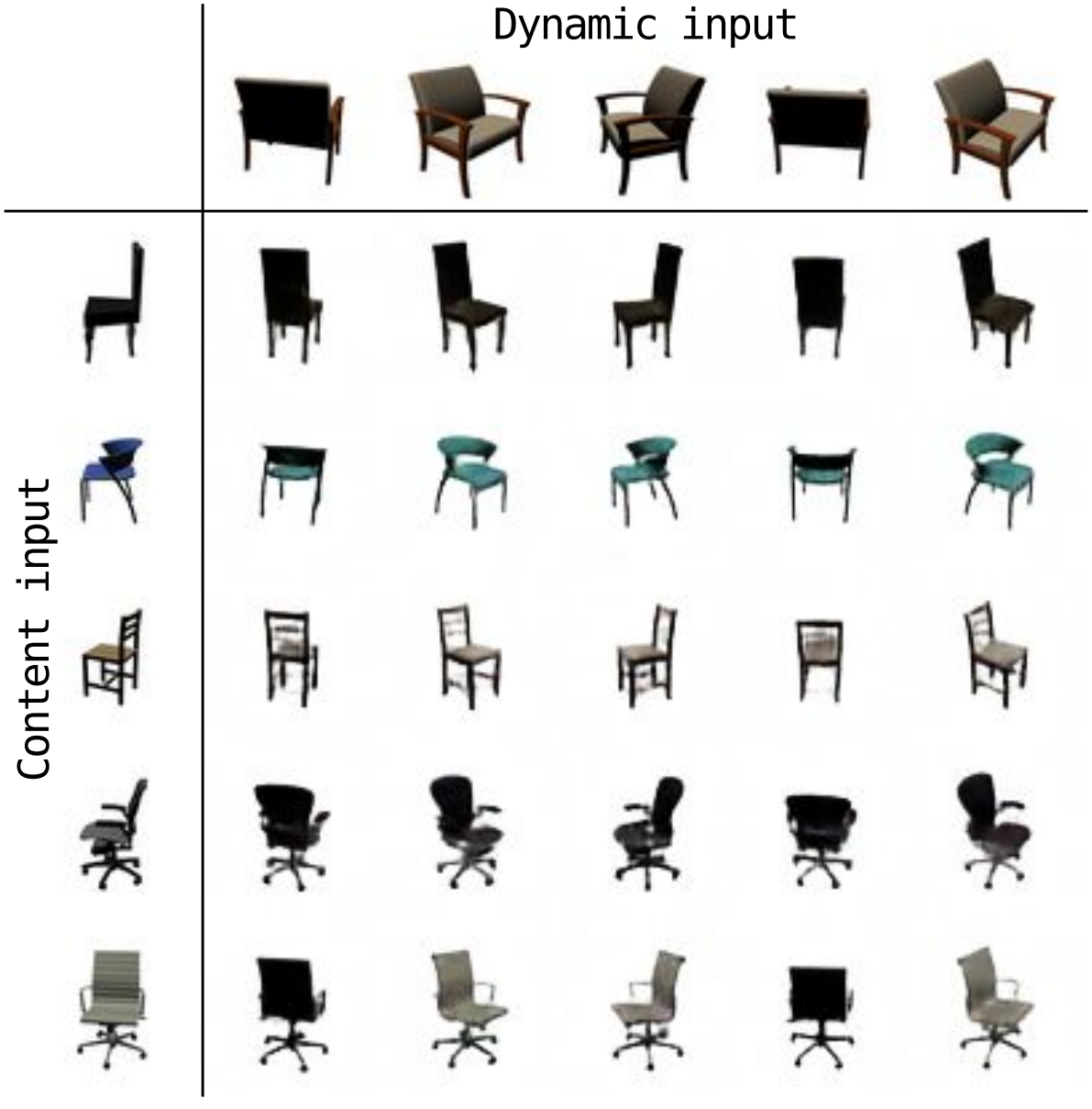
- From a strict S invariance constraint to a weaker one to take into account variations of observable content:

$$\frac{\partial E_S\big(V_\tau(t)\big)}{\partial t} = 0 \Rightarrow \mathcal{L}^S_{\text{reg}} = \Big\| E_S\big(V_\tau(t_0)\big) - E_S\big(V_\tau(t_1 - \tau)\big) \Big\|_2^2$$

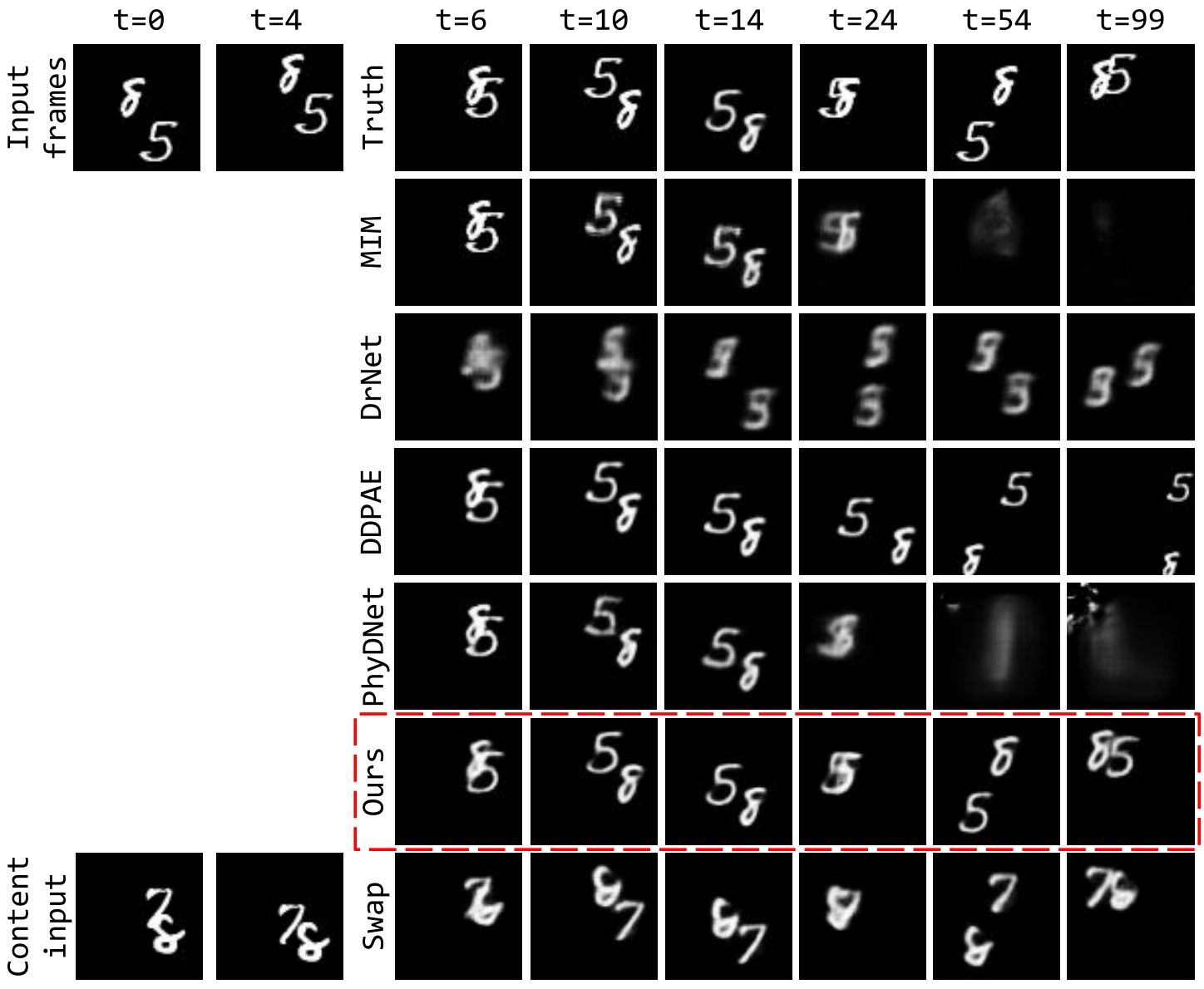
- Disentanglement loss:

$$\mathcal{L}^T_{\text{reg}} = \big\| T_{t_0} \big\|_2^2 = \Big\| E_T\big(V_\tau(t_0)\big) \Big\|_2^2$$

Experimental Results: Multiview & Prediction



Experimental Results: Multiview & Prediction



Experimental Results: Multiview & Prediction

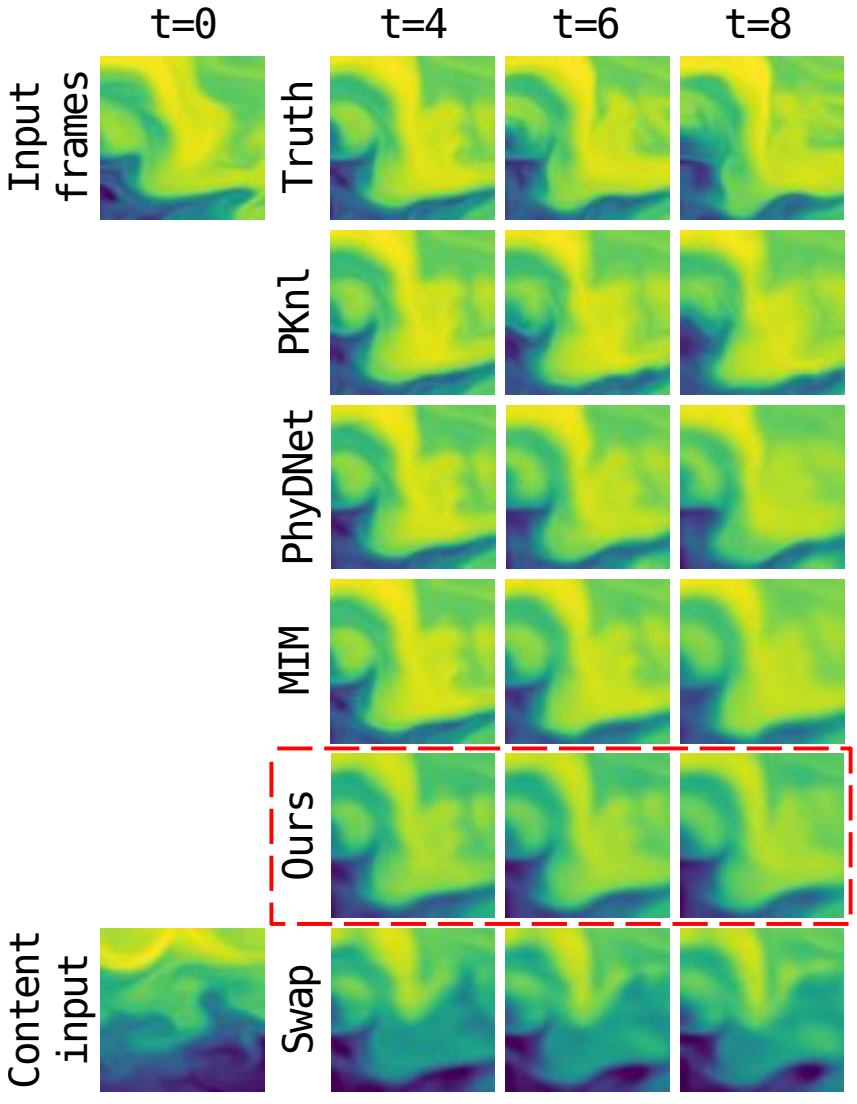


Table 1: Forecasting performance on WaveEq-100, WaveEq and SST of compared models with respect to indicated prediction horizons. Bold scores indicate the best performing method.

Models	WaveEq-100	WaveEq	SST			
	MSE		MSE		SSIM	
	$t+40$	$t+40$	$t+6$	$t+10$	$t+6$	$t+10$
PKnl	—	—	1.28	2.03	0.6686	0.5844
PhyDNet	—	—	1.27	1.91	0.5782	0.4645
SVG	—	—	1.51	2.06	0.6259	0.5595
MIM	—	—	0.91	1.45	0.7406	0.6525
Ours	$\mathbf{4.33 \times 10^{-5}}$	$\mathbf{1.44 \times 10^{-4}}$	**0.86**	**1.43**	**0.7466**	**0.6577**
Ours (without S)	1.33×10^{-4}	5.09×10^{-4}	0.95	1.50	0.7204	0.6446

Table 2: Prediction and content swap scores of all compared models on Moving MNIST. Bold scores indicate the best performing method.

Models	Pred. $(t+10)$		Pred. $(t+95)$		Swap $(t+10)$		Swap $(t+95)$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SVG	18.18	0.8329	12.85	0.6185	—	—	—	—
MIM	**24.16**	0.9113	16.50	0.6529	—	—	—	—
DrNet	14.94	0.6596	12.91	0.5379	14.12	0.6206	12.80	0.5306
DDPAE	21.17	0.8814	13.56	0.6446	**18.44**	0.8256	13.25	0.6378
PhyDNet	23.12	**0.9128**	16.46	0.3878	12.04	0.5572	13.49	0.2839
Ours	21.70	0.9088	**17.50**	**0.7990**	18.42	**0.8368**	**16.50**	**0.7713**

State-Space Predictive Models: Conclusion

- We study the relevance of state-space models for spatiotemporal data, with a focus on:
  - stochastic generation for videos;
  - variable separation for physical phenomena.
- Such models allow us to simulatenously perform accurate prediction and learn meaningful representations.
- Perspectives
  - Stochastic Differential Equations (e.g. Brownian Motion)
  - Neural ODE Formalism (invertible transformations?)
  - Adaptative Steps of Noise Introduction (i.e., asynchronously with frame rate)

Further Resources: code, models, animated samples

- `https://sites.google.com/view/srvp/`
- `https://github.com/JeremDona/spatiotemporal_variable_separation`

References

Babaeizadeh, Mohammad et al. (2018). "Stochastic Variational Video Prediction". In: *International Conference on Learning Representations*.

Denton, Emily and Vighnesh Birodkar (2017). "Unsupervised Learning of Disentangled Representations from Video". In: *Advances in Neural Information Processing Systems*. Ed. by Isabelle Guyon et al. Vol. 30. Curran Associates, Inc., pp. 4417–4426.

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*. Ed. by Daniel D. Lee et al. Vol. 29. 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. Vol. 30. Curran Associates, Inc., pp. 3604–3613.

Gregor, Karol et al. (2019). "Temporal Difference Variational Auto-Encoder". In: *International Conference on Learning Representations*.

Hsieh, Jun-Ting et al. (2018). "Learning to Decompose and Disentangle Representations for Video Prediction". In: *Advances in Neural Information Processing Systems*. Ed. by Samy Bengio et al. Vol. 31. Curran Associates, Inc., pp. 515–524.

Minderer, Matthias et al. (2019). "Unsupervised learning of object structure and dynamics from videos". In: *Advances in Neural Information Processing Systems*. Ed. by Hanna Wallach et al. Vol. 32. Curran Associates, Inc., pp. 92–102.

Villegas, Ruben et al. (2017). "Decomposing motion and content for natural video sequence prediction". In: *International Conference on Learning Representations*.

Yingzhen, Li and Stephan Mandt (July 2018). "Disentangled Sequential Autoencoder". 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. 5670–5679.