Parallelizing Autoregressive Generation with Variational State Space Models

Gaspard Lambrechts, Yann Claes, Pierre Geurts and Damien Ernst

Contributions

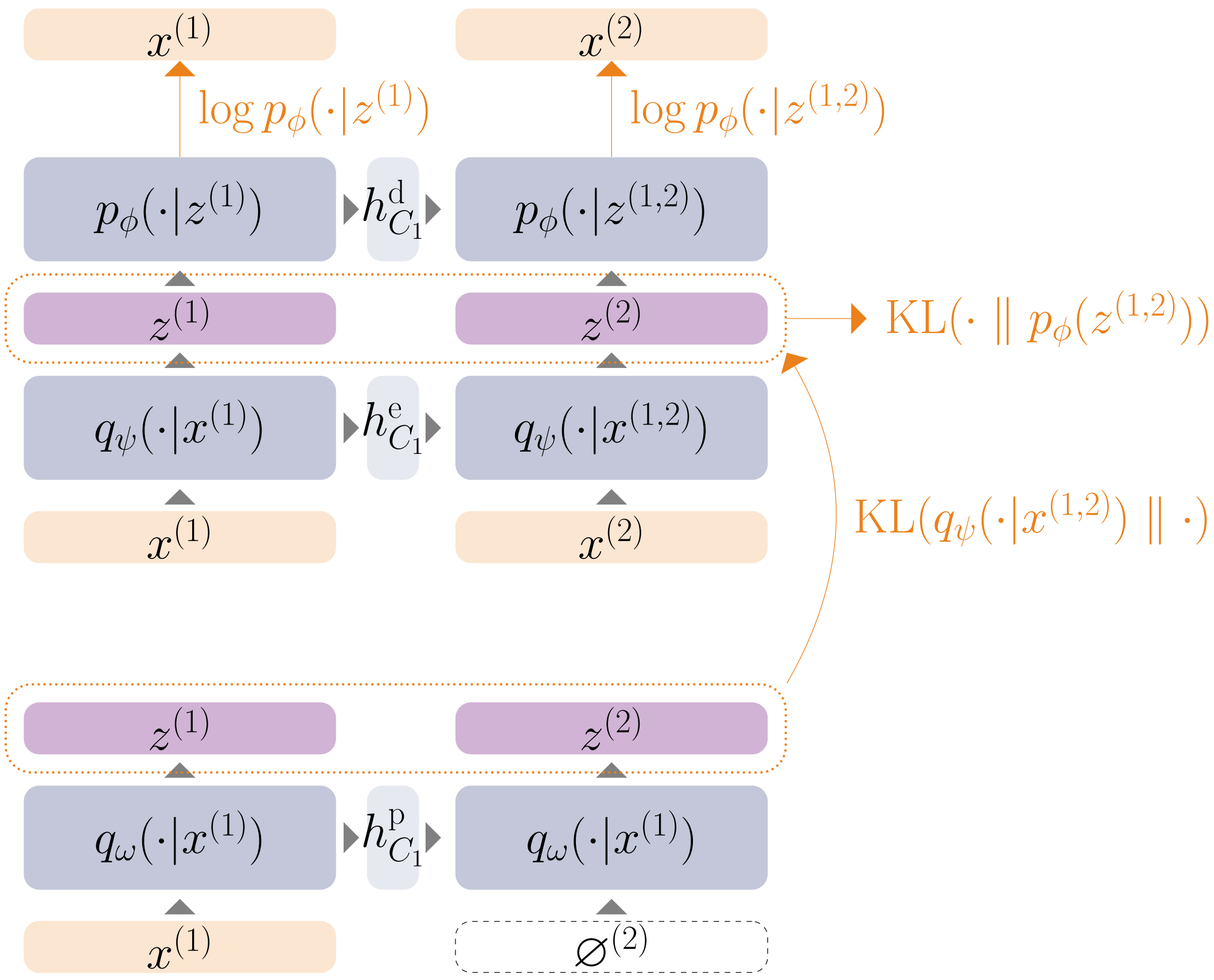
Summary

1. New VAE architecture with SSM encoder and decoder, enabling both **parallel training and generation**.
2. Introduction of a partial encoder based on an SSM allows **conditioning on partial realizations** (i.e., prompts).
3. Recurrence of the partial encoder and decoder allows **resuming generation without reprocessing** inputs.

Model	Generation	//	Prompt	Resume
Transformer	$O(T^2)$	✗	✓	✗
RNN	$O(T)$	✗	✓	✓
SSM	$O(T)$	✗	✓	✓
TVAE	$O(T^2)$	✓	✓	✗
VRNN	$O(T)$	✗	✓	✓
VSSM	$O(T)$	✓	✓	✓

Table 1. Model properties.

Training



1. Train posterior $q_\psi(z_{1:T}|x_{1:T})$ and generative $p_\phi(x_{1:T}|z_{1:T})$ according to the evidence lower bound (ELBO):

$$\max_{\phi,\psi} \mathop{\mathbb{E}}_{p(x_{1:T})} \underbrace{\left[\mathop{\mathbb{E}}_{q_\psi} \log p_\phi(x_{1:T}|z_{1:T}) - \mathrm{KL}(q_\psi(z_{1:T}|x_{1:T}) \parallel p_\phi(z_{1:T})) \right]}_{\mathrm{ELBO}_{\phi,\psi}(x_{1:T})}.$$

2. Train partial posterior $q_\omega(z_{1:T}|x_{1:C})$, for $C \sim \mathcal{U}([0,T])$:

$$\min_{\omega} \mathop{\mathbb{E}}_{p(x_{1:C})} \mathrm{KL}(p_\phi(z_{1:T}|x_{1:C}) \parallel q_\omega(z_{1:T}|x_{1:C})),$$

where,

$$p_\phi(z_{1:T}|x_{1:C}) = \mathop{\mathbb{E}}_{p(x_{1:T}|x_{1:C})} p_\phi(z_{1:T}|x_{1:T}), \approx \mathop{\mathbb{E}}_{p(x_{1:T}|x_{1:C})} q_\psi(z_{1:T}|x_{1:T}).$$

Generation

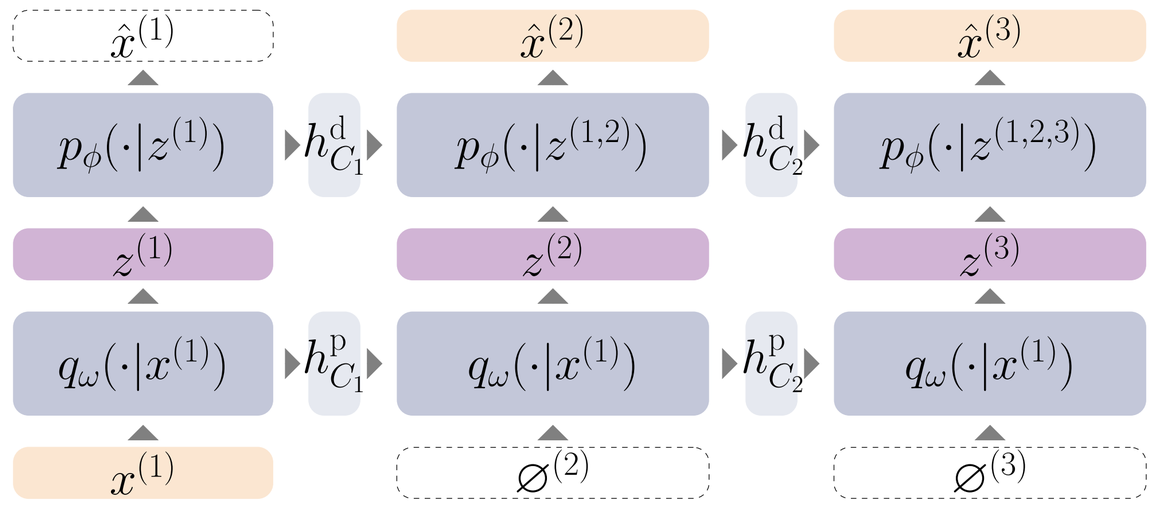


Fig 1. Given a prompt $\mathbf{x}^{(1)} = \mathbf{x}_{1:\mathbf{C}_1}$, generate $\hat{\mathbf{x}}^{(2)} = \hat{\mathbf{x}}_{\mathbf{C}_1:\mathbf{C}_2}$ in parallel, then resume generation to produce $\hat{\mathbf{x}}^{(3)} = \hat{\mathbf{x}}_{\mathbf{C}_2:\mathbf{C}_3}$.

Architecture

Variational State Space Model (VSSM)

- **Prior distribution** uniform on a discrete latent space:

$$p_\phi(z_{1:T}) = \prod_{t=1}^{T} p_\phi(z_t) = \prod_{t=1}^{T} \frac{1}{N^Z},$$

- **Generative distribution** based on an SSM f^{dec}_ϕ:

$$p_\phi(x_{1:T}|z_{1:T}) = \prod_{t=1}^{T} \mathcal{P}(x_t|f^{\mathrm{dec}}_\phi(z_{1:t})),$$

- **Posterior distribution** based on an SSM f^{enc}_ψ:

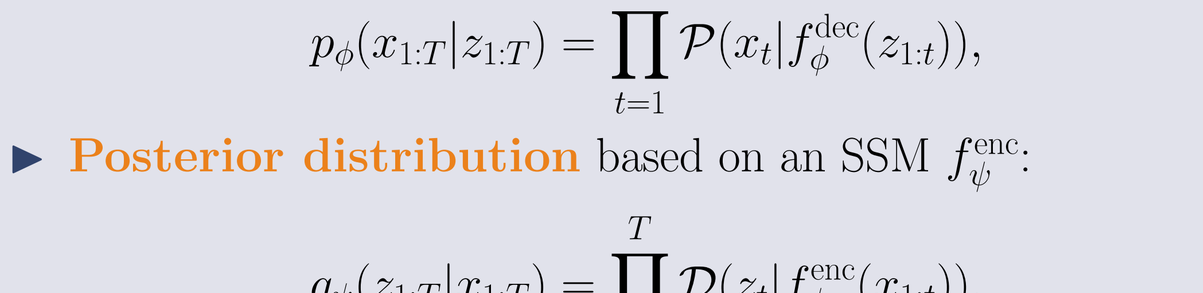
$$q_\psi(z_{1:T}|x_{1:T}) = \prod_{t=1}^{T} \mathcal{D}(z_t|f^{\mathrm{enc}}_\psi(x_{1:t})),$$

- **Partial posterior distribution** based on an SSM f^{par}_ω:

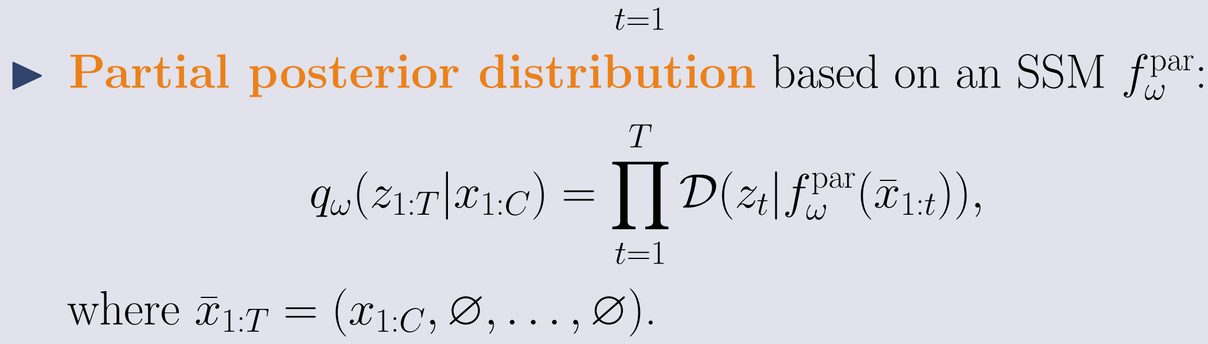
$$q_\omega(z_{1:T}|x_{1:C}) = \prod_{t=1}^{T} \mathcal{D}(z_t|f^{\mathrm{par}}_\omega(\bar{x}_{1:t})),$$

where $\bar{x}_{1:T} = (x_{1:C}, \varnothing, \ldots, \varnothing)$.

Results



(a) MNIST training, statistics and samples.



(b) CIFAR training, statistics and samples.

Fig 2. Validation likelihood (full and partial), test statistics (likelihood and time), samples (unconditional and prompted).