Uncertainty inspired solutions to stochastic problems via variational autoencoders

Fatemeh Saleh

Australian National University

Outline

- Introduction to Variational Autoencoders

- Recent Applications:
  - A Stochastic Conditioning Scheme for Diverse Human Motion Prediction
  - UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders

- Conclusion

Variational Autoencoders

Generative Models

- Informally:
  - A model that can generate new data after learning from the dataset.
- More formally:
  - A generative model models the joint distribution $P(X, Y)$ of the observation X and the target Y.
  - A discriminative model models the conditional distribution $P(Y|X)$.

Discriminative versus Generative

- Discriminative Model
  - Tries to learn the discriminative information from the data
  - Example: Classify C1 vs C2 vs C3
    - Finds a good decision boundary by directly modeling conditional distribution $P(Y|X)$
    - Learns mappings from inputs to classes

- Generative Model
  - Tries to learn the distribution of the data
  - Models the distribution of inputs characteristic of the class
  - For classification, builds a model of $P(X|Y)$ and then applies Bayes Rule

Generative Models

- Given a training set of examples, e.g., images of cats:

$$x_i \sim P_{data}$$
$$i = 1,2,\ldots,n$$

- We want to learn a probability distribution $p(x)$ over images x for
  - **Generation:** If we sample $x_{new} \sim p(x)$, x_{new} should look like a cat (sampling)
  - **Density estimation:** $p(x)$ should be high if x looks like a cat, and low otherwise
  - **Unsupervised representation learning:** The model should be able to learn what these images have in common, e.g., ears, tail, etc. (features)

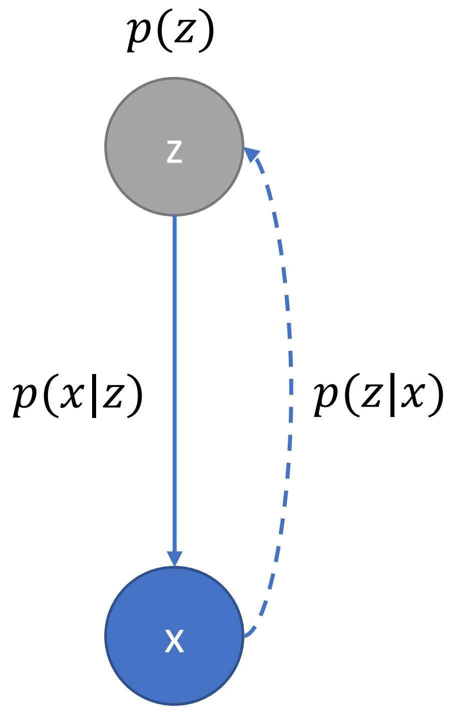
Latent Variable Models

- LVM defines a distribution over observations x by using a latent variable z and specifying
  - The prior distribution $p(z)$ for the latent variable
  - The likelihood $p(x|z)$ that connects the latent variable to the observation

- The joint distribution

$$p(x, z) = p(x|z)p(z)$$

- We are interested in computing the marginal likelihood $p(x)$ and the posterior distribution $p(z|x)$



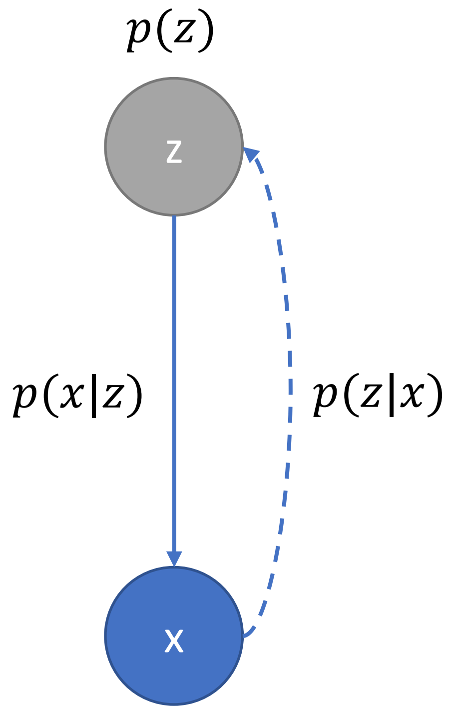
Why latent variables?

- Latent variables explains the data
- To generate an observation from such explanation

$$z \sim p(z)$$
$$x \sim p(x|z)$$

- The inverse of generation is called inference,

$$z \sim p(z|x)$$



Inference

- Why inference is important?
  - Explaining the observation
    - Inferring the posterior distribution for a datapoint allows us to determine what latent configurations could have plausibly generate such datapoint.
  - Learning
    - Training latent variable models requires performing the inference.

Inference

- Exact inference is hard!

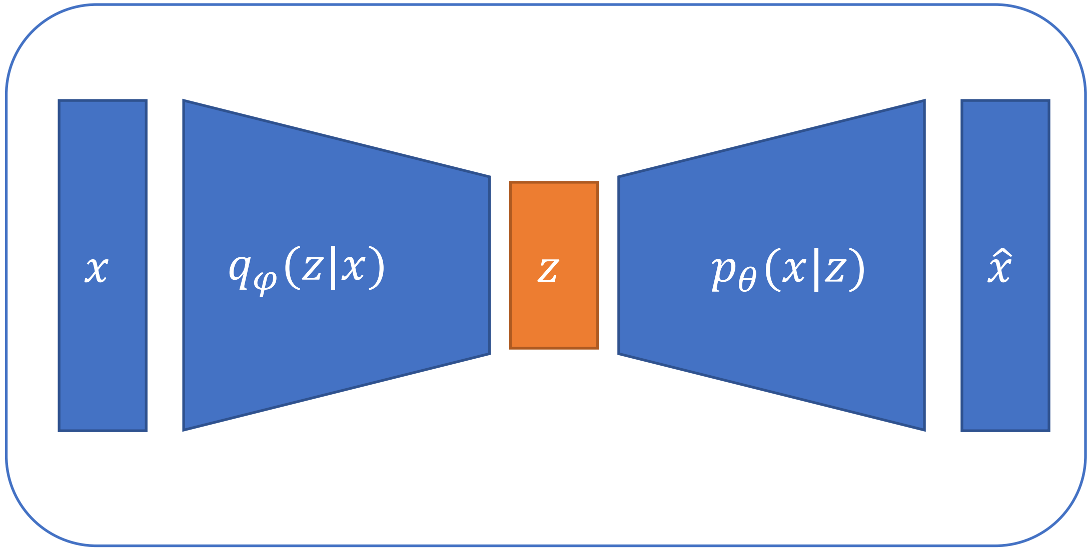
$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(x,z)}{\int p(x,z)dz}$$

- Computing $\int p(x,z)dz$ is intractable as it requires considering all configurations of z for a reasonable approximation of $p(x)$.

There are some exceptions!

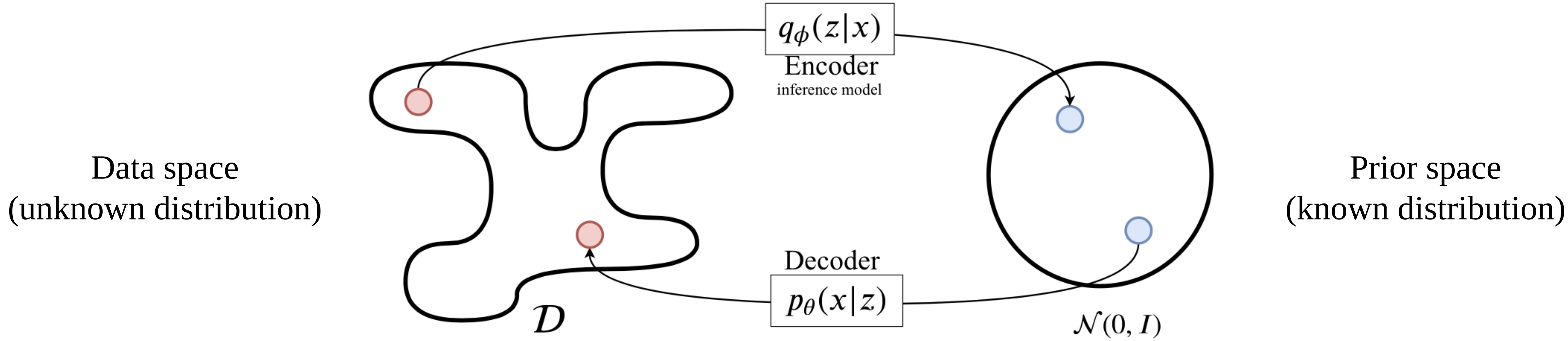
Inference

- One way to avoid interactable inference is to approximate posterior distribution



Variational Autoencoders

- An overview



Variational Autoencoders

- VAE's objective function
  - $q_\varphi(z|x) \approx p_\theta(z|x)$

$$D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right] = \mathbf{E}_{z\sim q_\phi(z|x)}\left[\log\frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$$

Background on KL Divergence

- KL divergence provides a way of quantifying the difference between two distributions.
- KL divergence is defined as

$$KL(q(z)||p(z)) = \mathrm{E}_{q(z)}\left[\log \frac{q(z)}{p(z)}\right]$$

- KL divergence is
  - Non-negative
  - KL=0 if $q(z) = p(z)$
  - Non-symmetric, $KL(q(z)||p(z)) \neq KL(p(z)||q(z))$

Variational Autoencoders

- Expanding the KL divergence between the approximate and true posterior distributions

$$D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right] = \mathbf{E}_{z\sim q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$$

$$= \mathbf{E}_{z\sim q_\phi(z|x)}\left[\log q_\phi(z|x) - \boxed{\log p_\theta(z|x)}\right]$$

True posterior

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)}$$

Variational Autoencoders

- The data distribution $p_\theta(x)$ is independent of z, so

$$D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right] = \mathbf{E}_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(x|z) - \log p(z)\Big] + \log p_\theta(x)$$

- By putting $p_\theta(x)$ to the left side

$$D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right] - \log p_\theta(x) = \mathbf{E}_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(x|z) - \log p(z)\Big]$$

$$\log p_\theta(x) - D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right] = \mathbf{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z) - \big(\log q_\phi(z|x) - \log p(z)\big)\Big]$$

$$= \mathbf{E}_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - \boxed{\mathbf{E}_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p(z)\Big]}$$

Variational Autoencoders

- So, it can be written as

$$\underbrace{\log p_\theta(x)}_{\text{Log-likelihood of the data}} - \underbrace{D_{KL}\left[q_\phi(z|x)||p_\theta(z|x)\right]}_{\text{KL divergence between the approximate and the true posterior}} = \underbrace{\mathbf{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{Reconstruction loss}} - \underbrace{D_{KL}\left[q_\phi(z|x)||p(z)\right]}_{\text{KL divergence between the approximate posterior and prior}}$$

not computable!

According to definition, it is positive!

- Thus, we can only optimize the lower bound on the data log-likelihood

$$\log p_\theta(x) \geq \mathbf{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left[q_\phi(z|x)||p(z)\right]$$

Variational Autoencoders

- Therefore, the ELBO of the VAE is

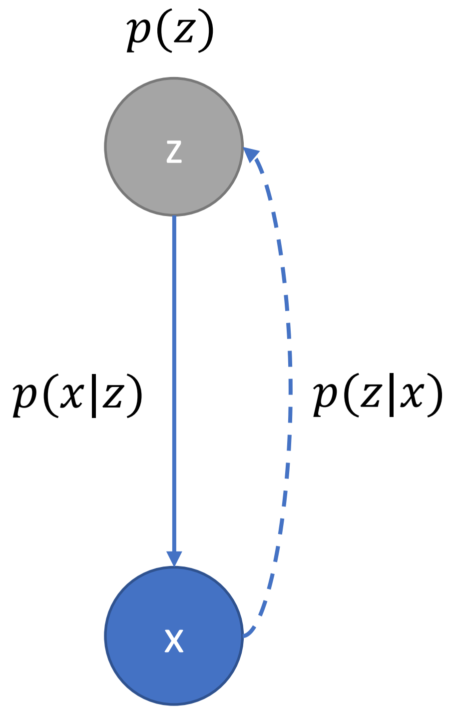
$$\log p_\theta(x) \geq \mathbf{E}_{z\sim q_\phi(z|x)}\big[\log p_\theta(x|z)\big] - D_{KL}\big[q_\phi(z|x)||p(z)\big]$$

- And, similarly, the ELBO of conditional VAE is

$$\log p_\theta(x|c) \geq \mathbf{E}_{z\sim q_\phi(z|x)}\big[\log p_\theta(x|z,c)\big] - D_{KL}\big[q_\phi(z|x,c)||p(z|c)\big]$$

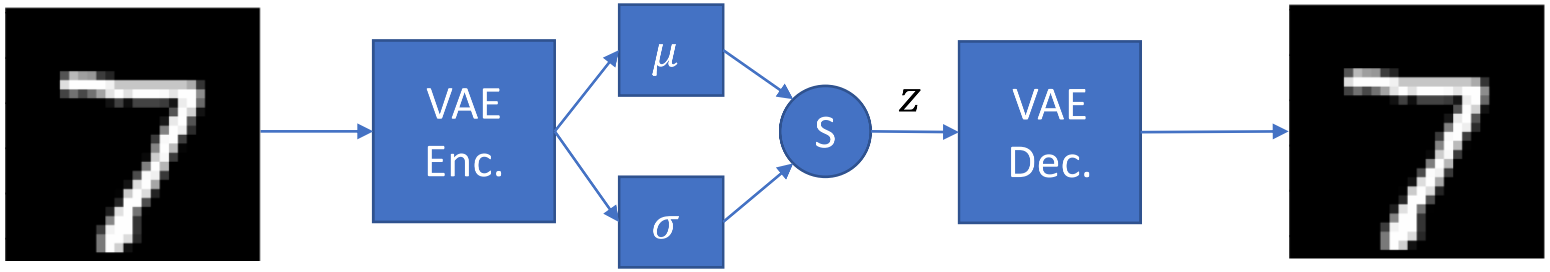
Variational Autoencoders

- In practice,
  - Prior: $p(z) = \mathcal{N}(0, I)$
  - Encoder: $q_\varphi(z|x) = \mathcal{N}\left(\mathrm{Net}_\varphi(x), \mathrm{diag}\left(\mathrm{Net}_\varphi(x)\right)\right)$
  - Decoder: $p_\theta(x|z) = \mathrm{Net}_\theta(z)$



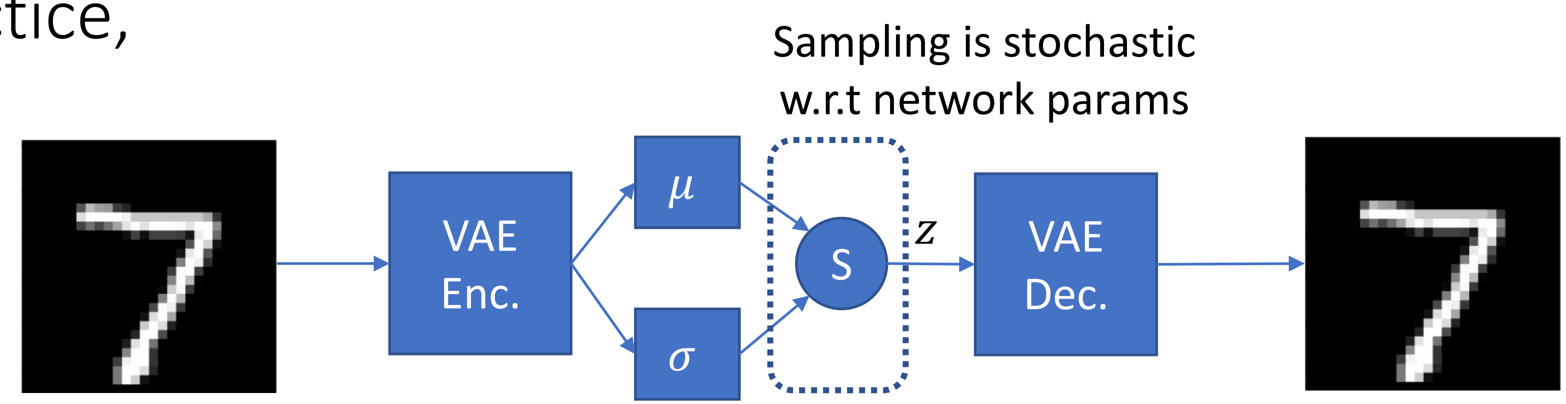
Variational Autoencoders

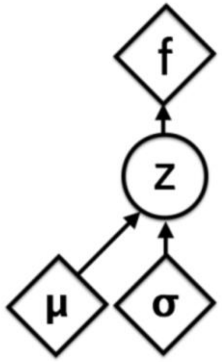
- In practice,



Variational Autoencoders

- In practice,

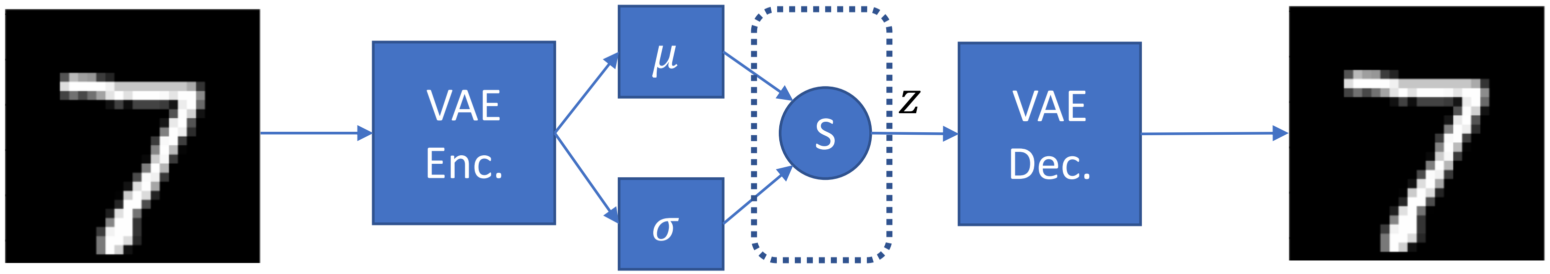


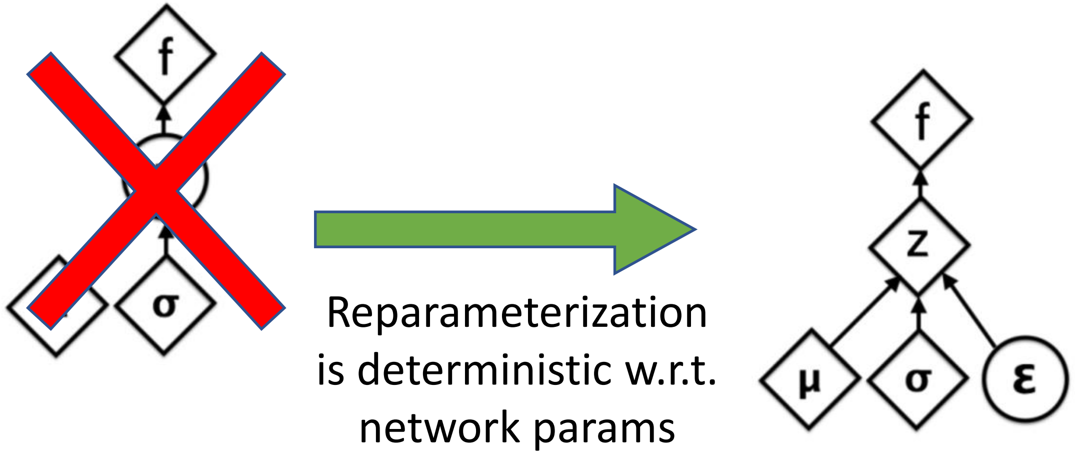


Variational Autoencoders

- In practice,

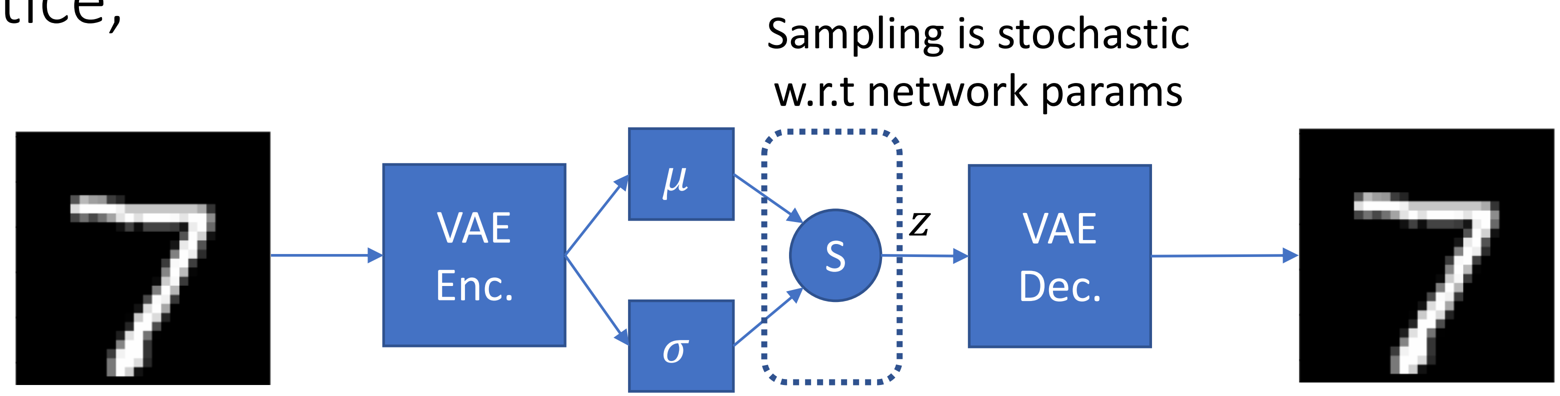
Sampling is stochastic w.r.t network params



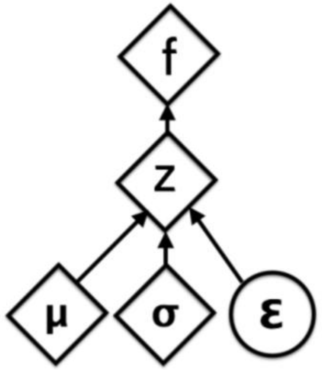


Variational Autoencoders

- In practice,

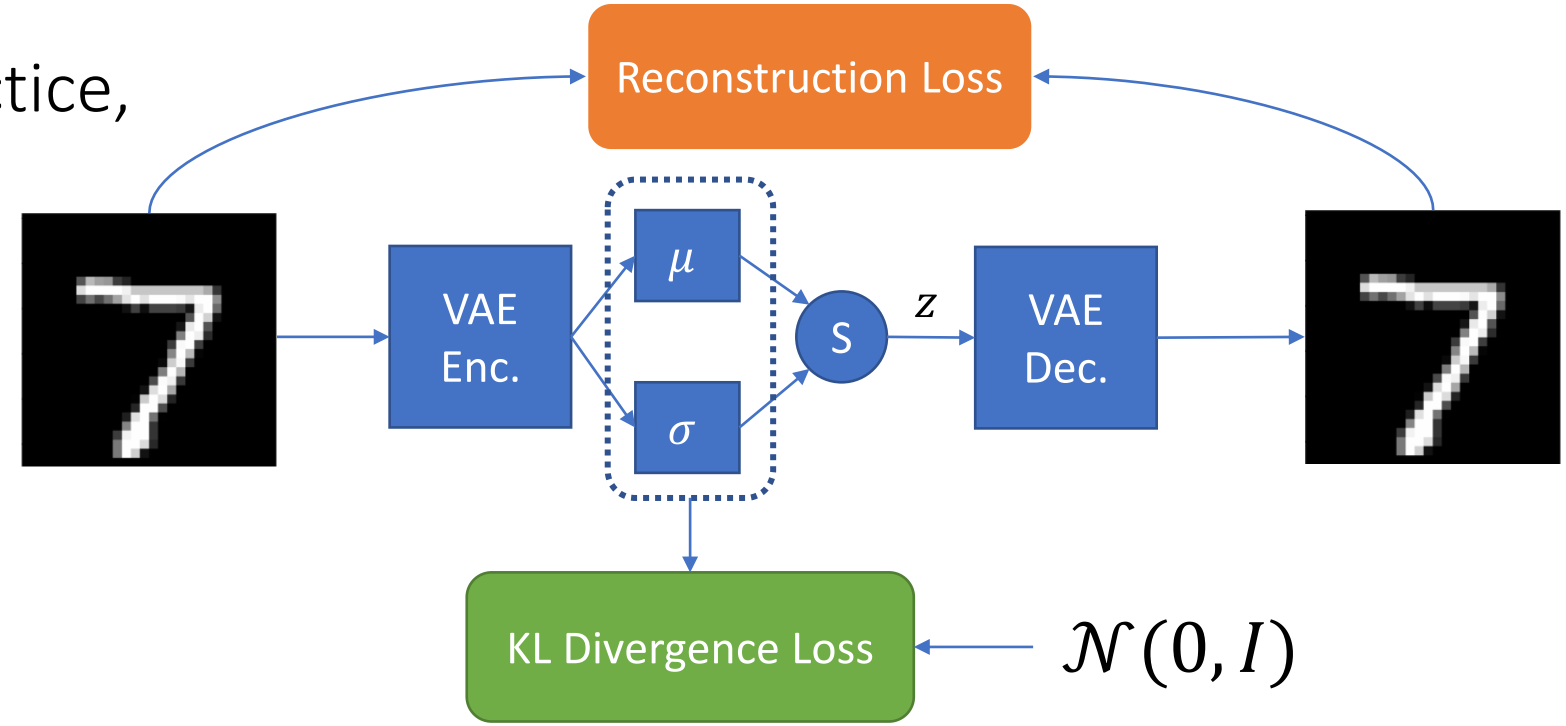


- To sample from posterior
  - Sample epsilon from $\epsilon \sim \mathcal{N}(0, I)$
  - Shift and scale ϵ given estimated posterior parameters
  - $z = \mu + \sigma \odot \epsilon$
  - This is equivalent to $\mathrm{z} \sim \mathcal{N}(\mu, \sigma)$



Variational Autoencoders

- In practice,



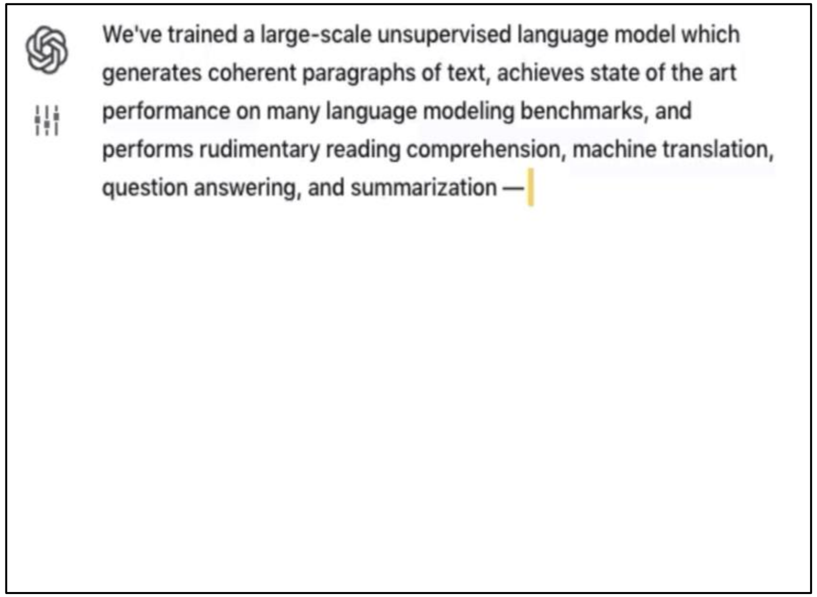
$$\log p_\theta(x) \geq \mathbf{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x|z) \right] - D_{KL}\left[q_\phi(z|x) || p(z) \right]$$

A Stochastic Conditioning Scheme for Diverse Human Motion Prediction

Sadegh Aliakbarian, Fatemeh Saleh, Mathieu Salzmann, Lars Petersson, Stephen Gould, CVPR 2020

Problem definition

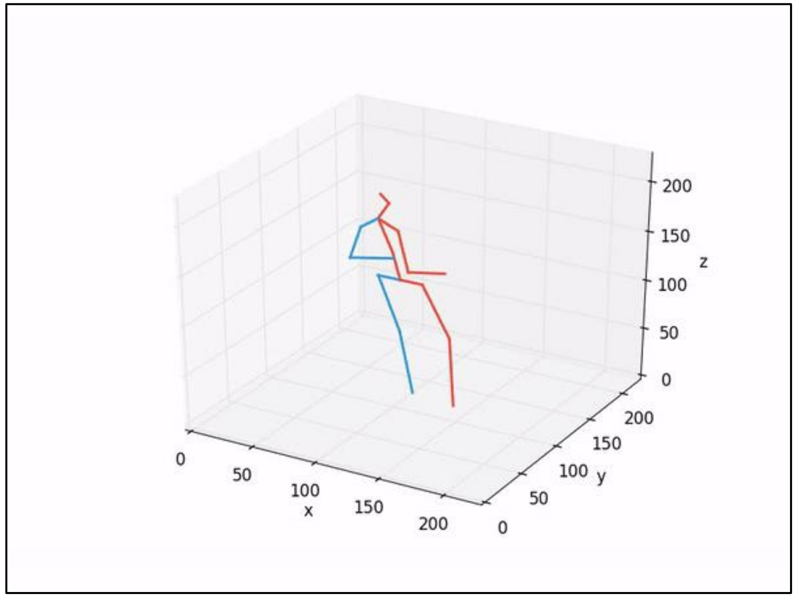
- Generating a sequence given a strong conditioning signal



Text completion

A red double-decker bus is parked at the bus stop.

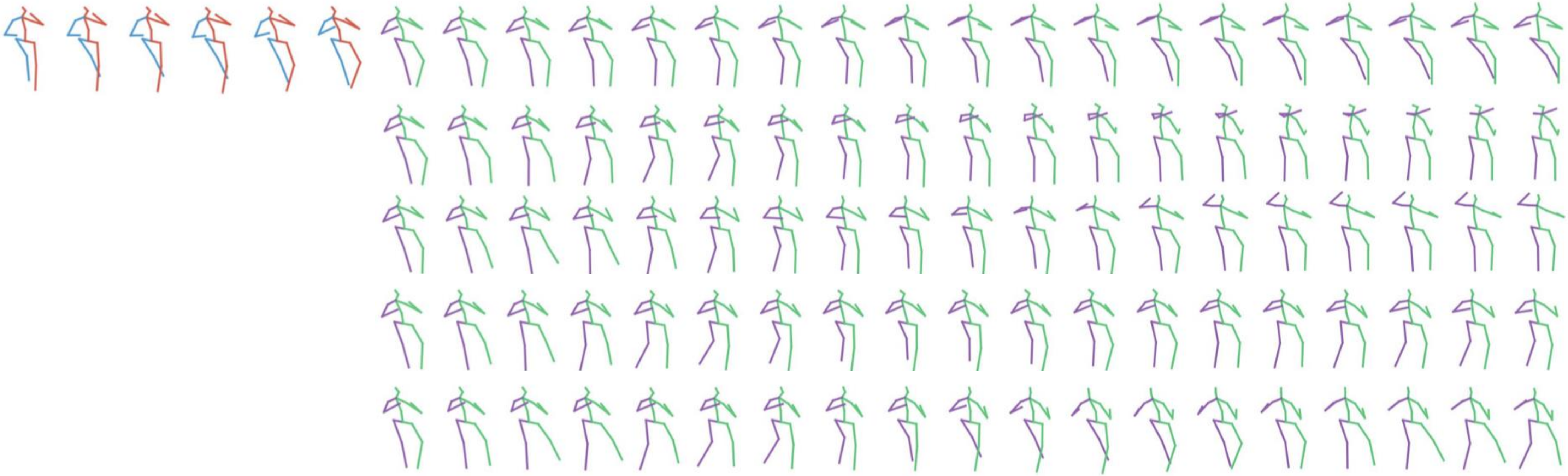
Image captioning



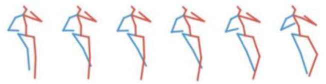
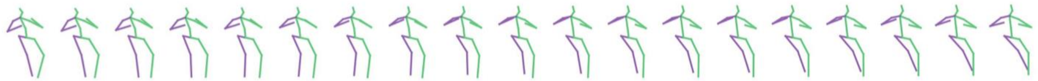
Human motion prediction

Problem definition

- Even given strong conditions, the solution might be ambiguous!



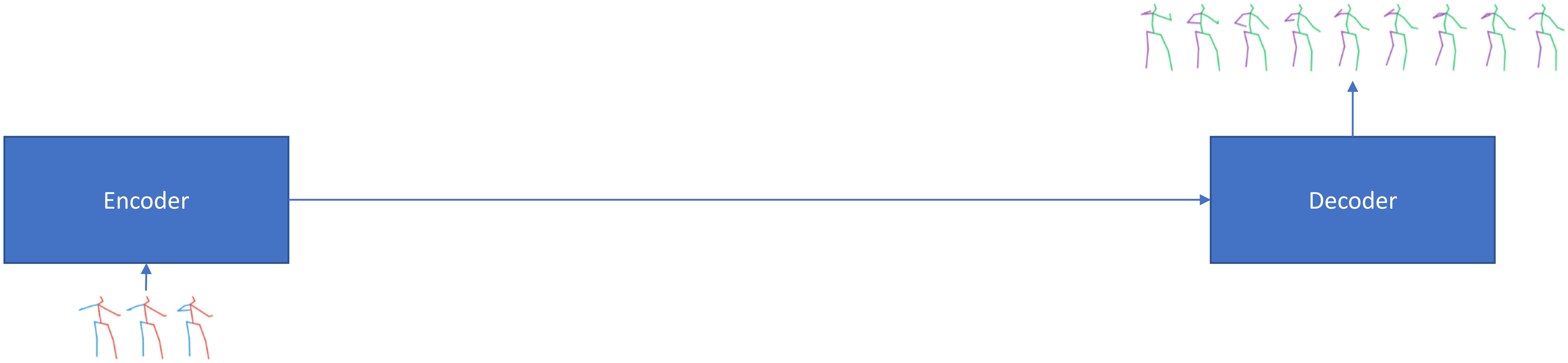
Problem definition

- Challenge
  - Related datasets are often deterministic
  - One sample per condition, e.g.:
  - Condition 
  - Sample: 

- Solution:
  - Learning the underlying distribution instead of a mapping.
  - We use VAEs

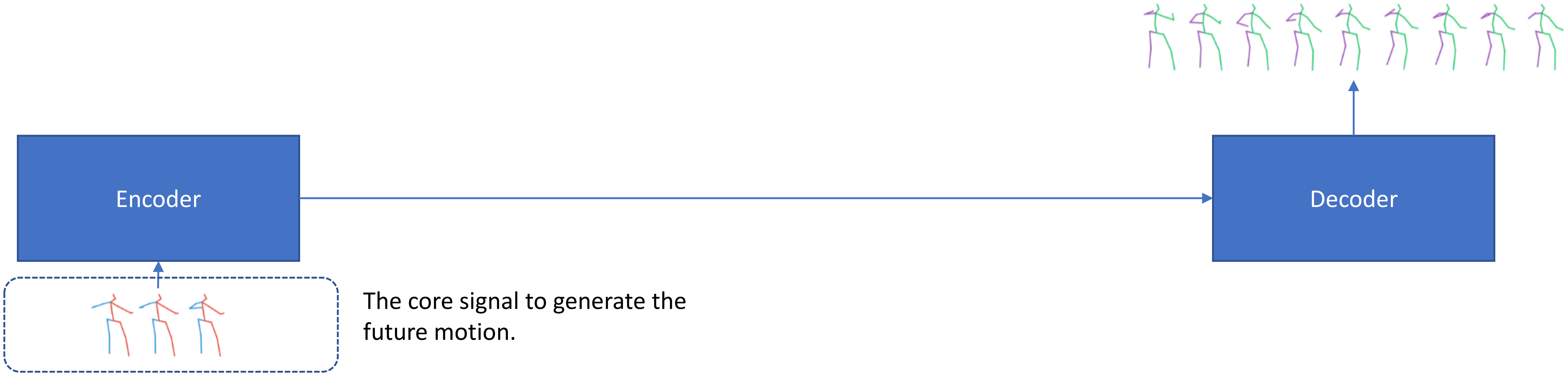
Human Motion Prediction

- Deterministic motion prediction (high level overview of the state-of-the-art)



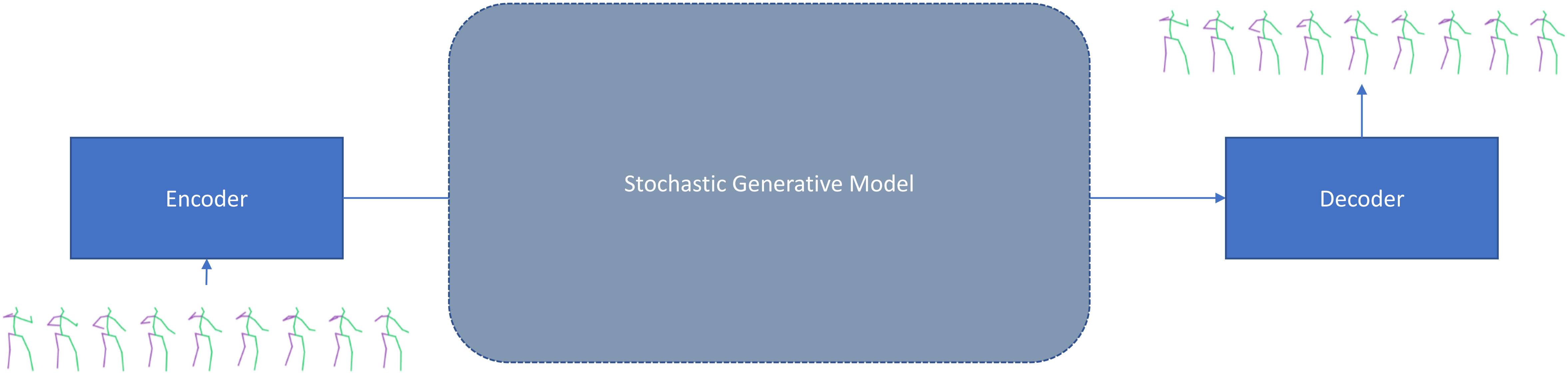
Human Motion Prediction

- Deterministic motion prediction (high level overview of the state-of-the-art)



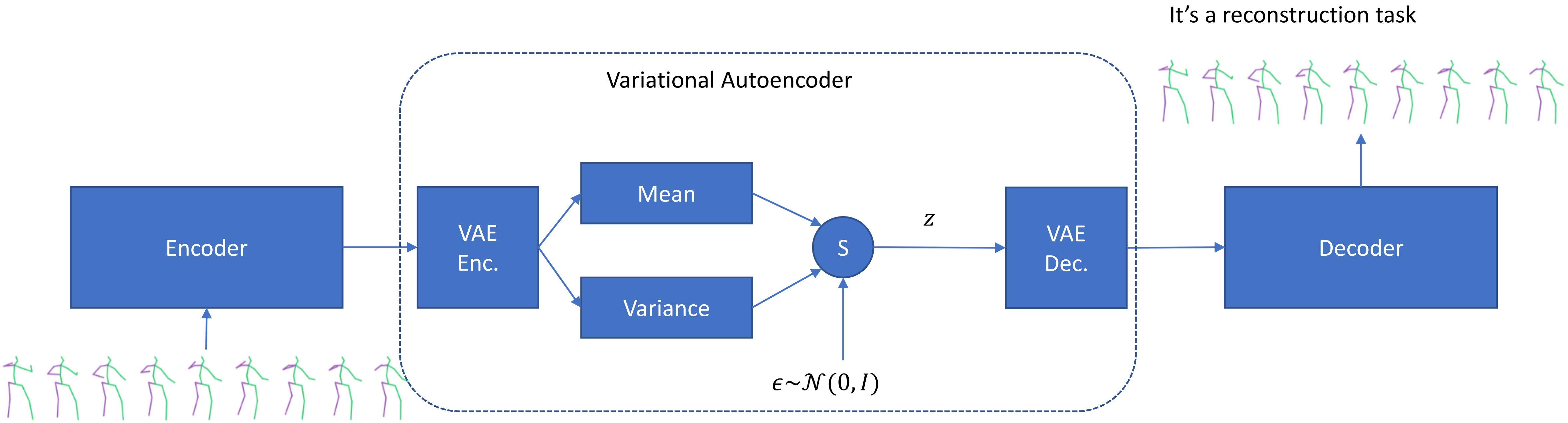
Human Motion Prediction

- From deterministic to stochastic

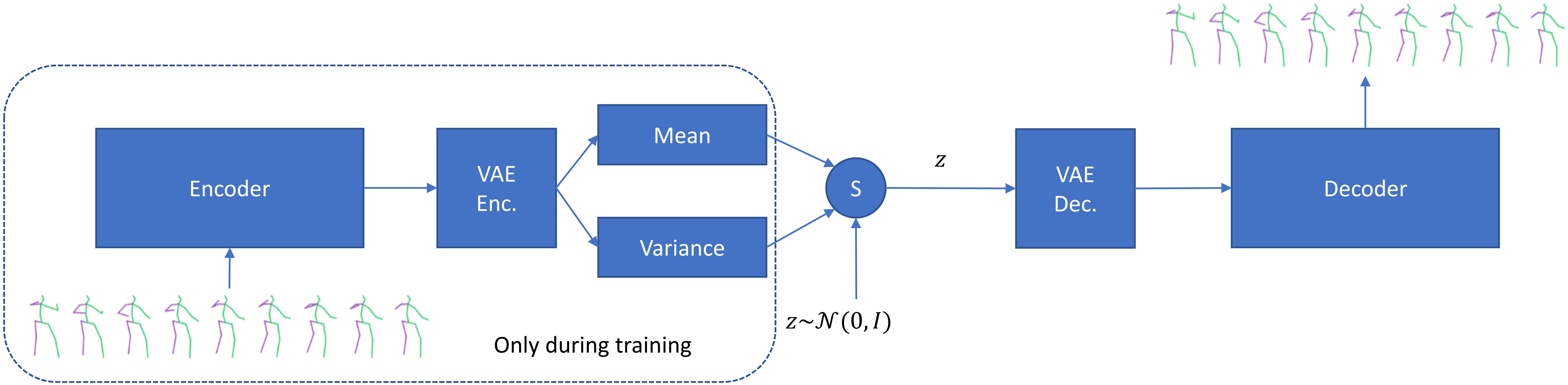


Human Motion Prediction

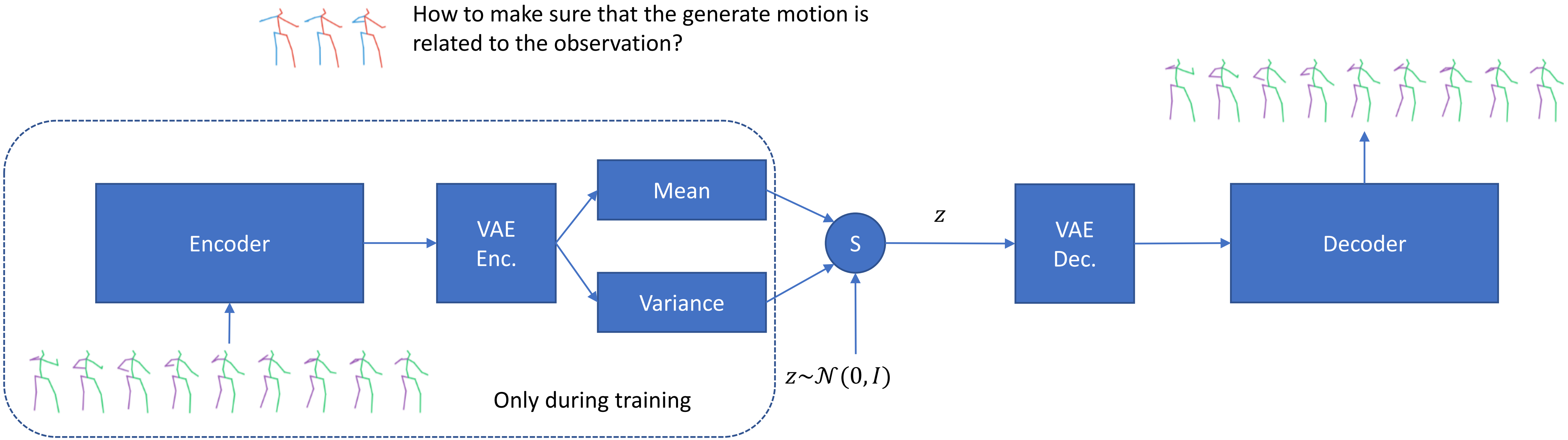
- Stochastic motion prediction baseline



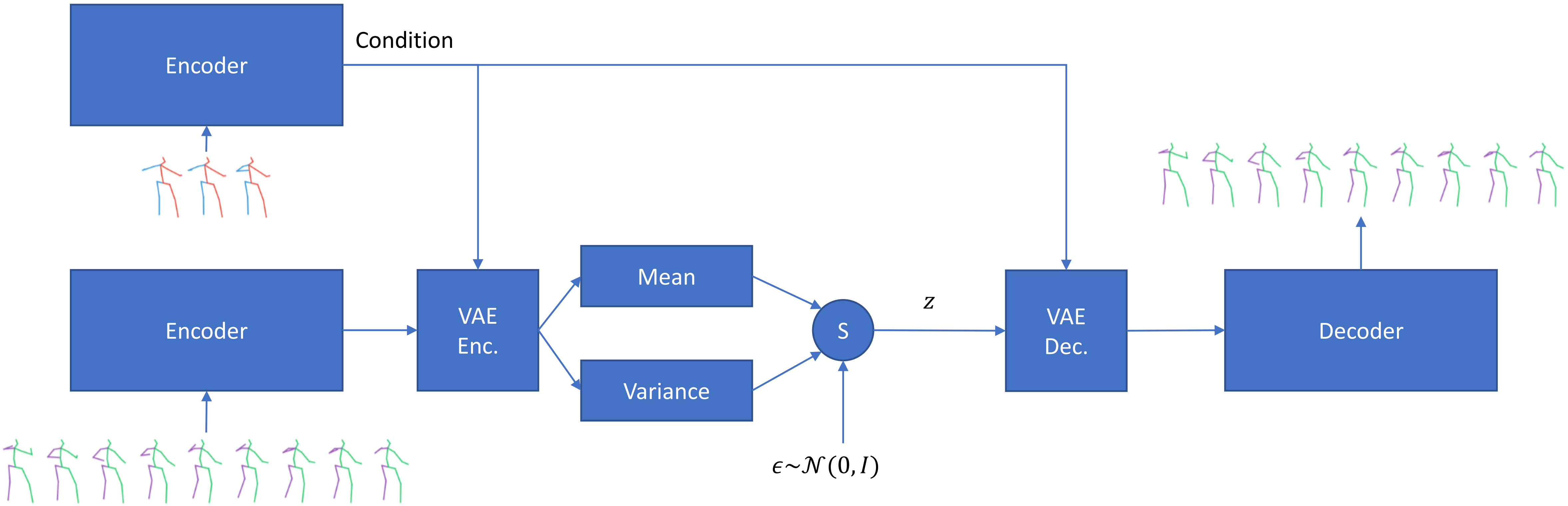
Human Motion Prediction



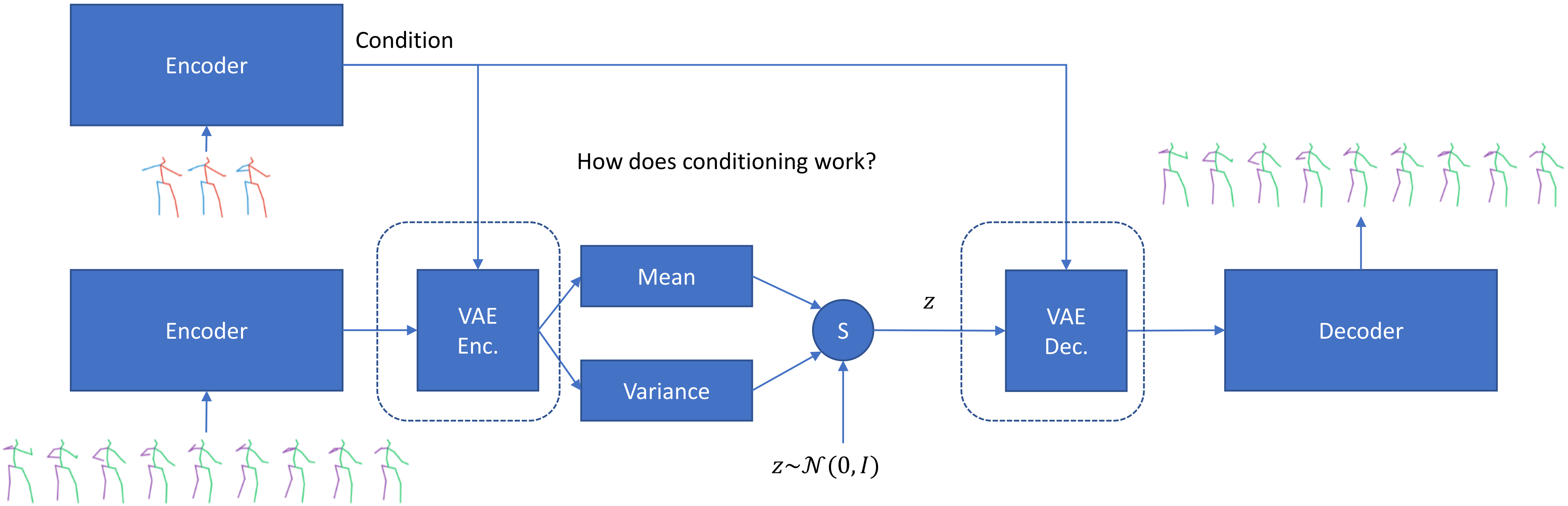
Human Motion Prediction



Human Motion Prediction



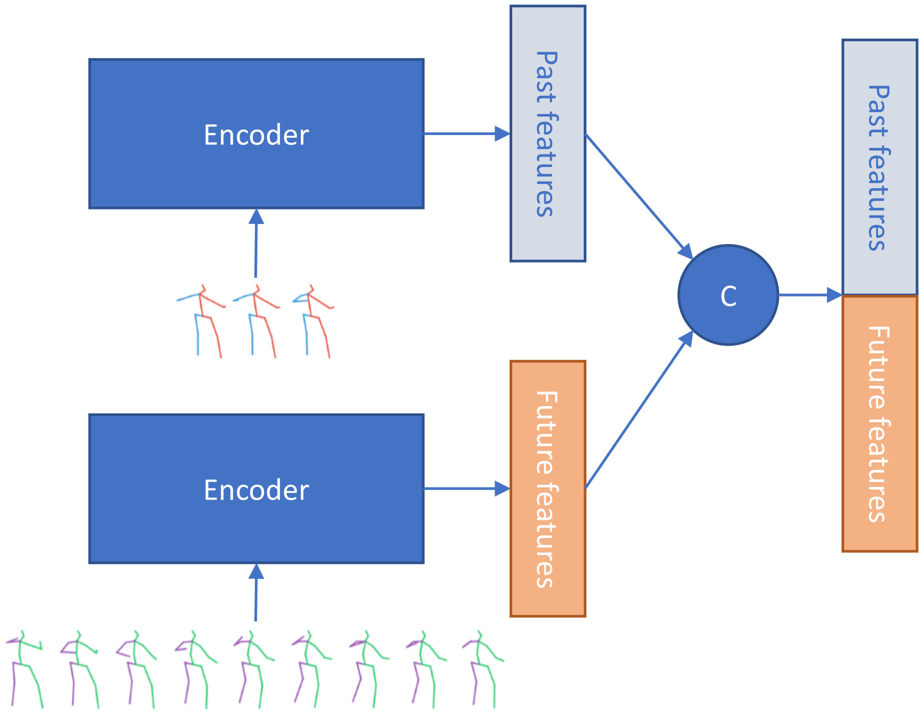
Human Motion Prediction



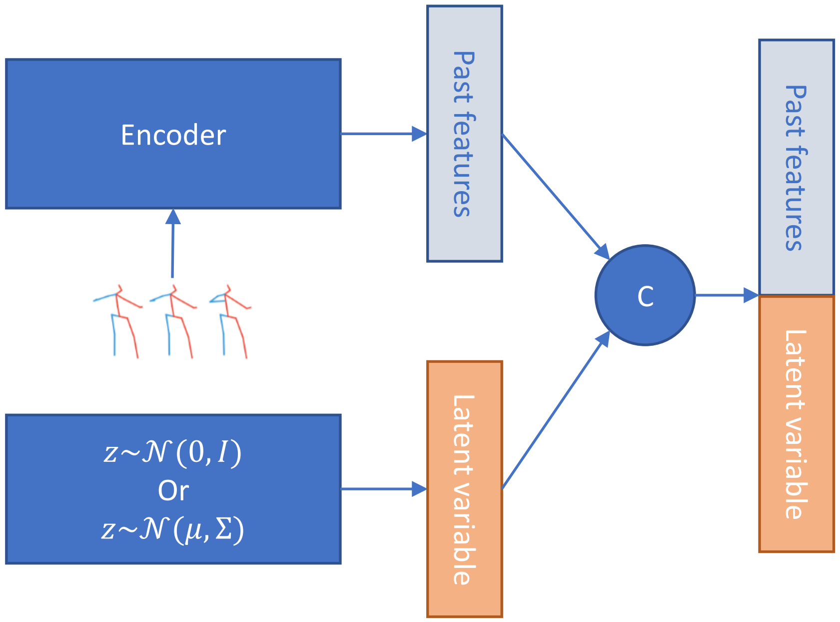
Human Motion Prediction

- How conditioning is done?
  - Deterministically, usually by concatenation

Conditioning the VAE Encoder



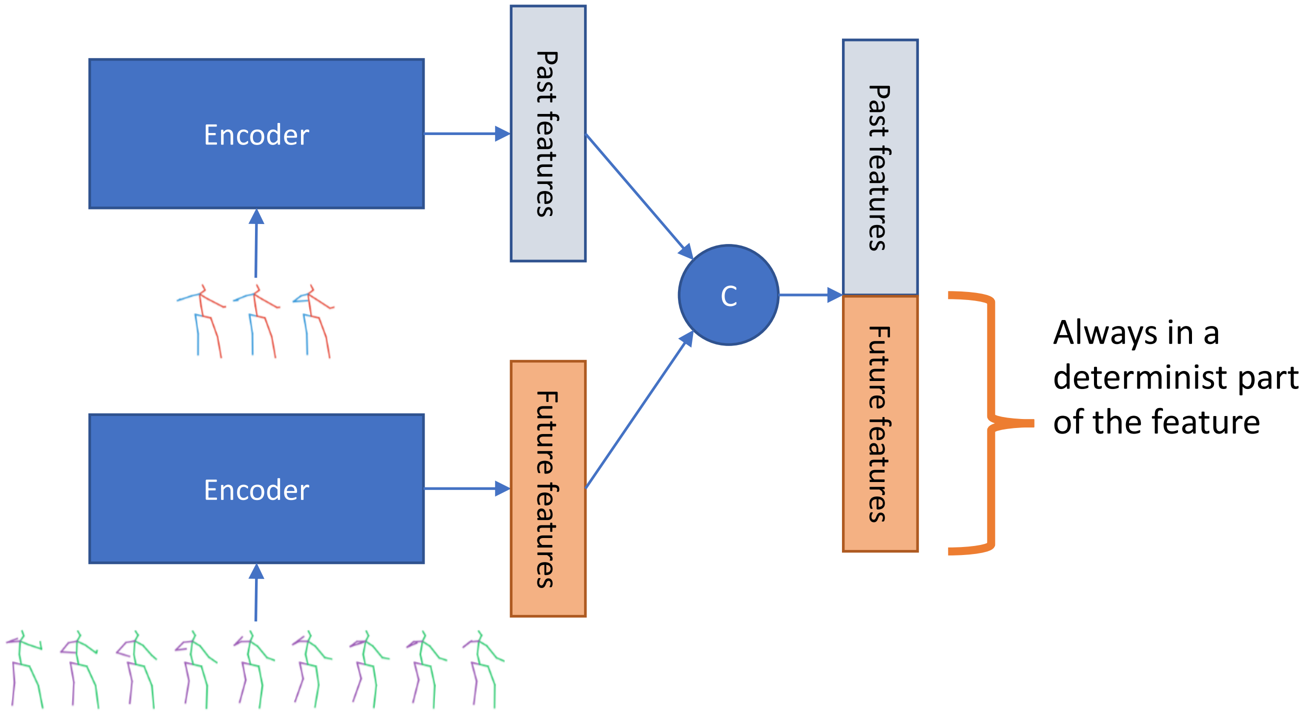
Conditioning the VAE Decoder



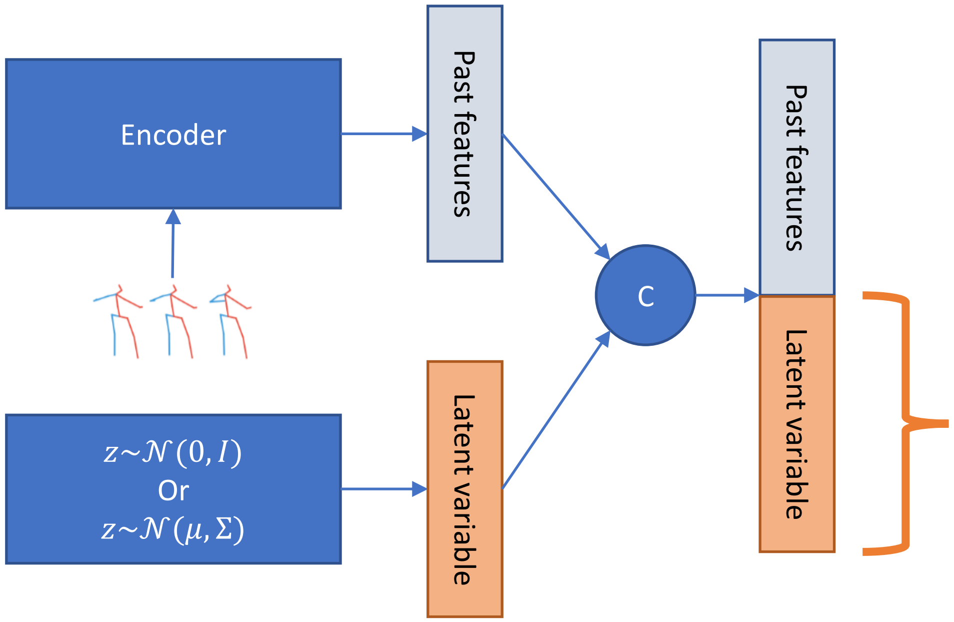
Human Motion Prediction

- How conditioning is done?
  - Deterministically, usually by concatenation

Conditioning the VAE Encoder

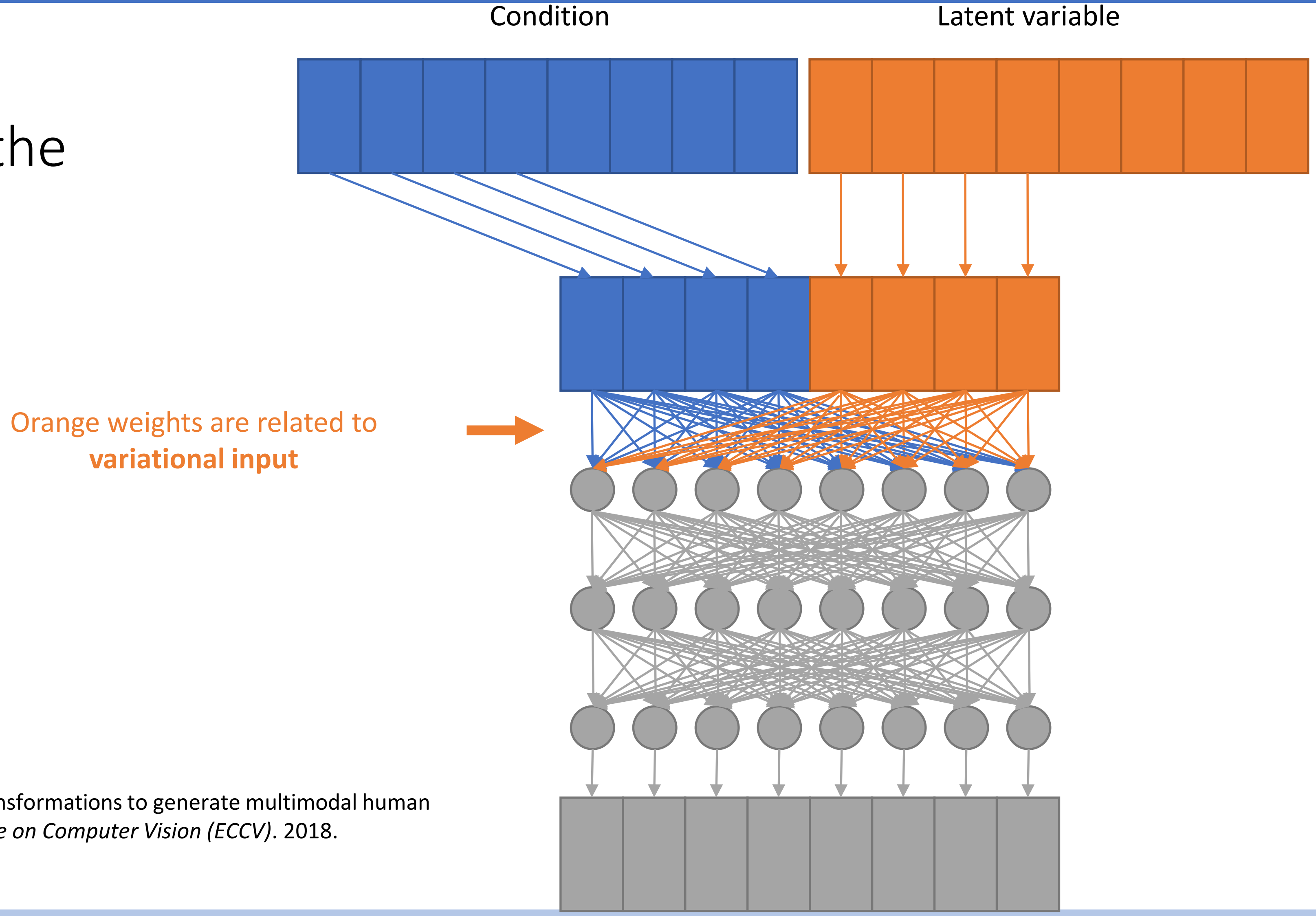


Conditioning the VAE Decoder



Human Motion Prediction

- E.g., conditioning the decoder
  - As in MT-VAE*



* Yan, Xinchen, et al. "Mt-vae: Learning motion transformations to generate multimodal human dynamics." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Human Motion Prediction

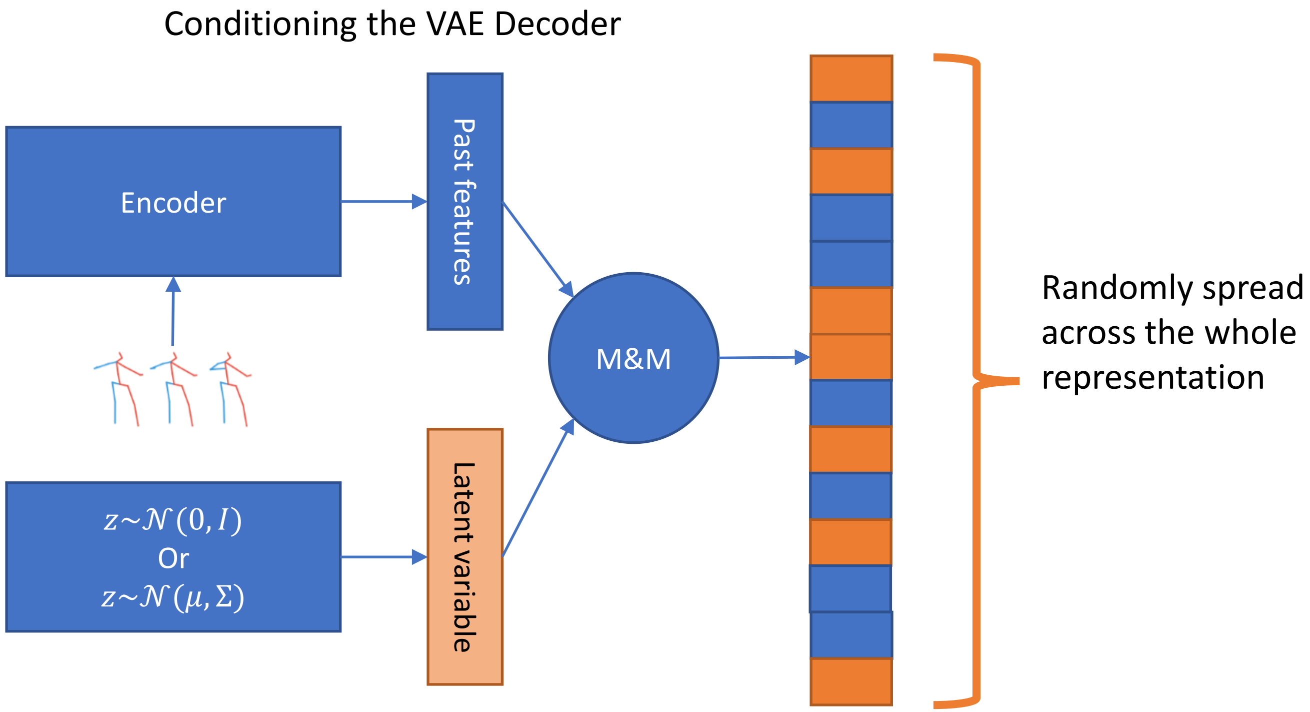
- Deterministic conditioning when dealing with strong conditions

- Posterior Collapse
  - Concatenation operation allows the decoder to learn to ignore the latent variable
  - Posterior collapses to the prior
  - Latent variable carries no information about the data

ignoring the latent variable $\approx$ ignoring the root of variation

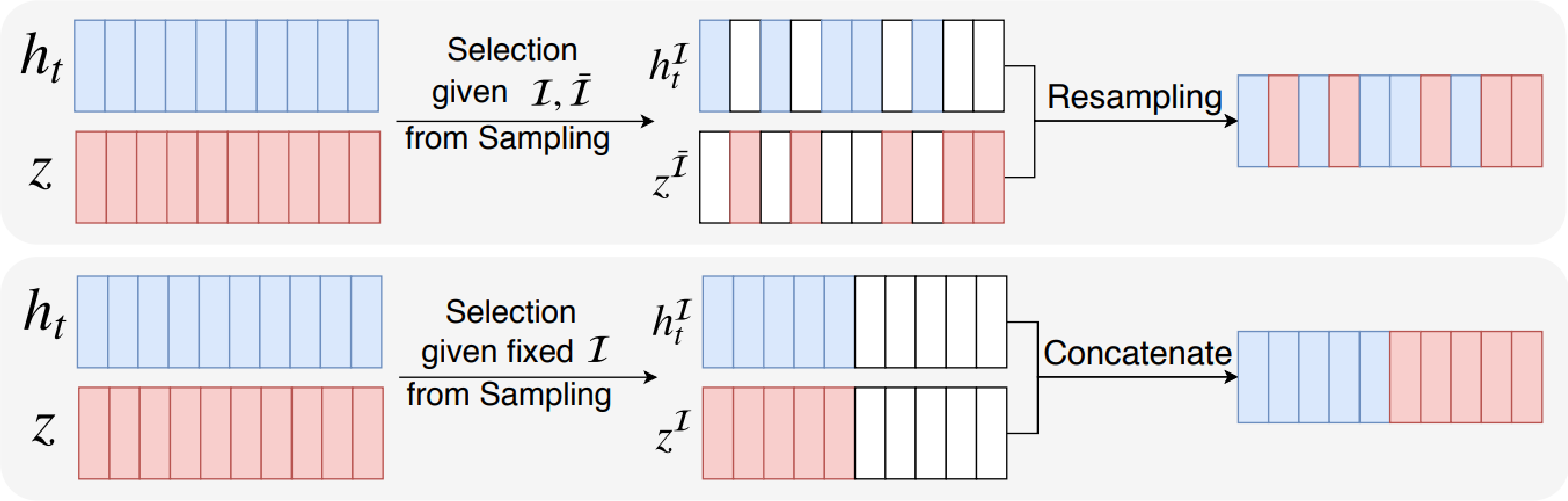
Mix-and-match Perturbation

- Our solution
  - Replacing the deterministic conditioning to a stochastic one



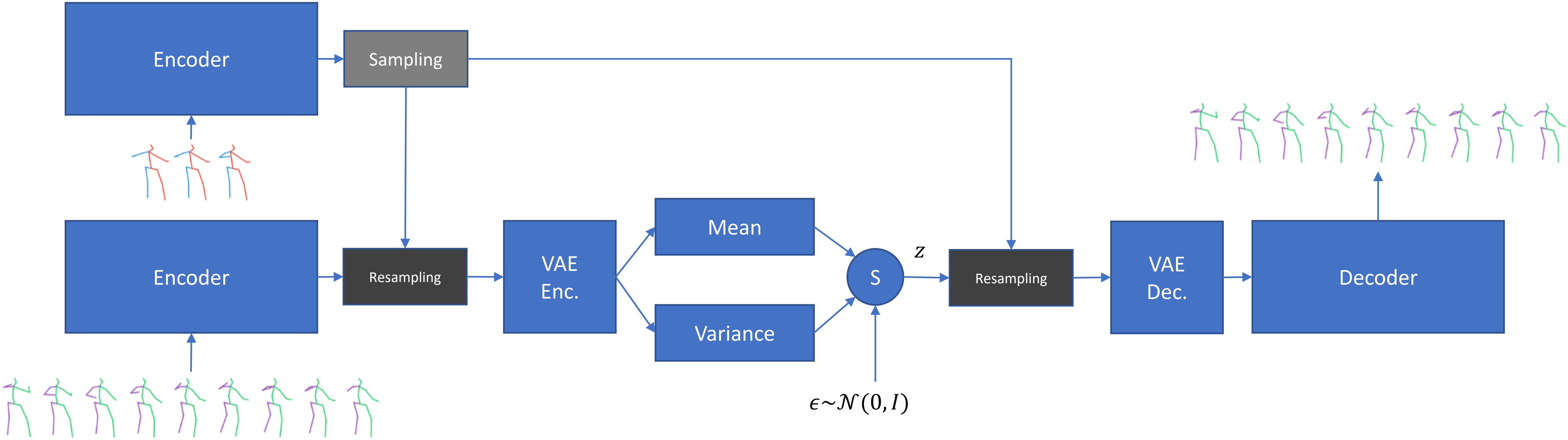
Mix-and-Match Perturbation

- Mix-and-Match versus concatenation



Mix-and-Match Perturbation

- An overview of the framework



Mix-and-Match Perturbation

- Learning: Loss functions

$$\mathcal{L}_{motion} = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{rot}(X_i) + \mathcal{L}_{skl}(X_i)\Big) + \lambda\mathcal{L}_{prior}$$

Experiments and Results

- Dataset: Human 3.6M dataset

Experiments and Results

- (Quantitative) Evaluation Metrics:
  - Human evaluations
  - Deterministic evaluations, i.e., against GT

- We propose two new evaluation metrics to quantitatively measure the quality and diversity of motions
  - Quality
  - Diversity
  - Diversity as a function of Quality

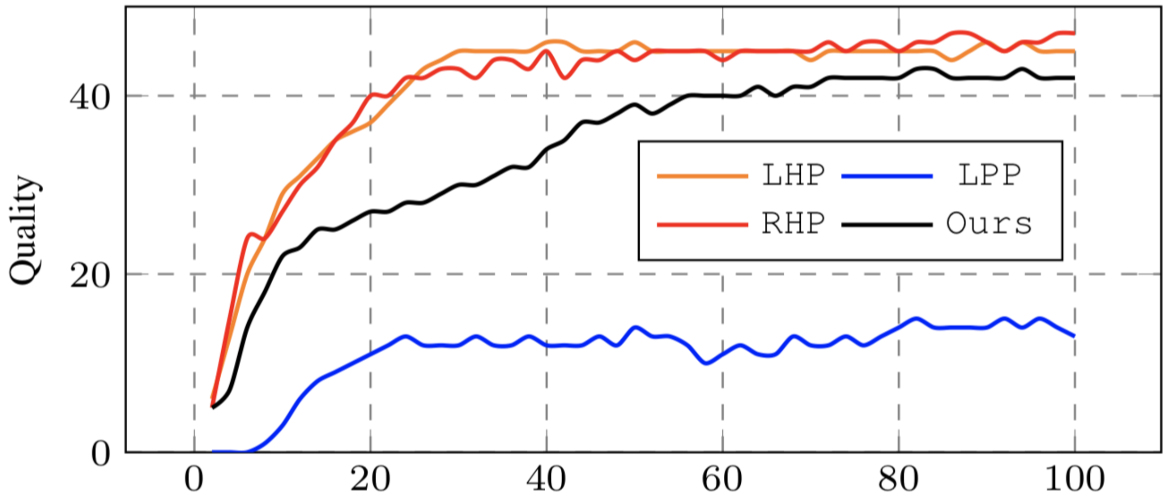
Experiments and Results

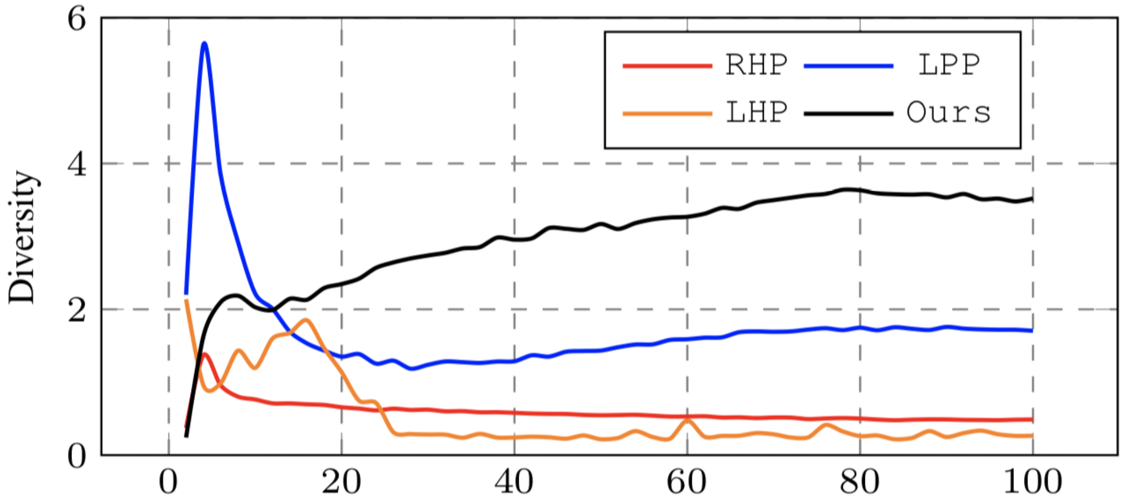
- Diversity Measure
  - make use of the average distance between all pairs of generated motions.

- Quality Measure
  - A binary classifier trained to discriminate real (ground-truth) samples from fake (generated) ones.
  - The accuracy of this classifier on the test set is inversely proportional to the quality of the generated motions.

Experiments and Results

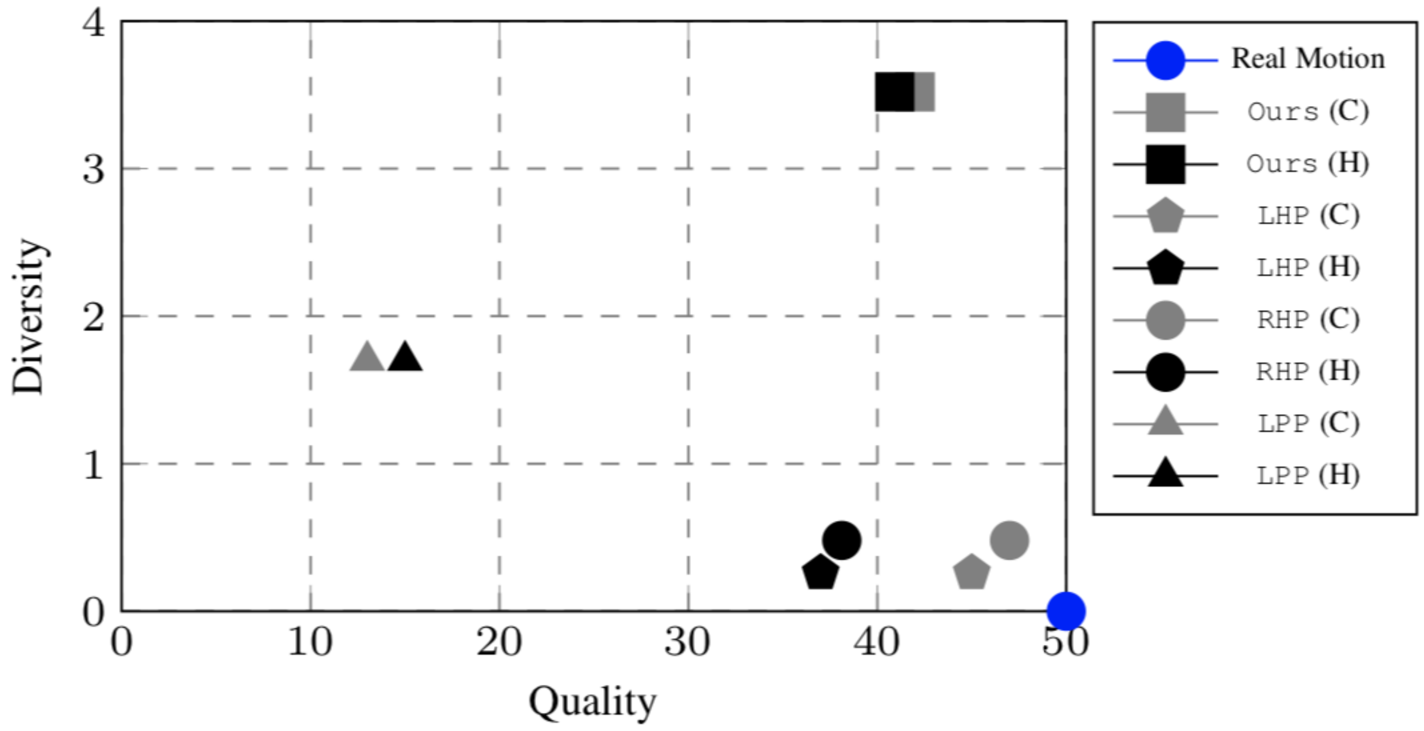
- Evaluating quality and diversity





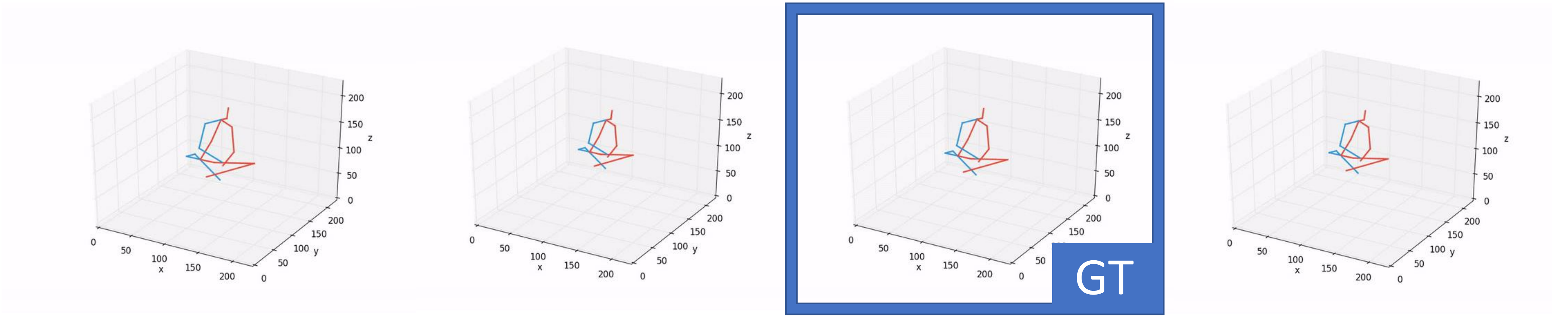
Experiments and Results

- Diversity as a function of quality



Experiments and Results

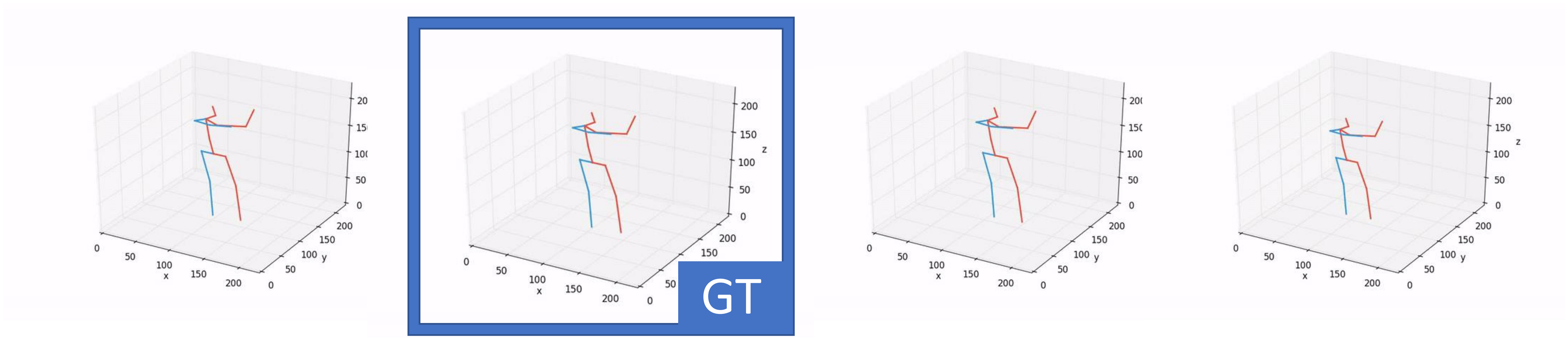
- Qualitative results



Observation: 16 frames
Generation: 60 frames

Experiments and Results

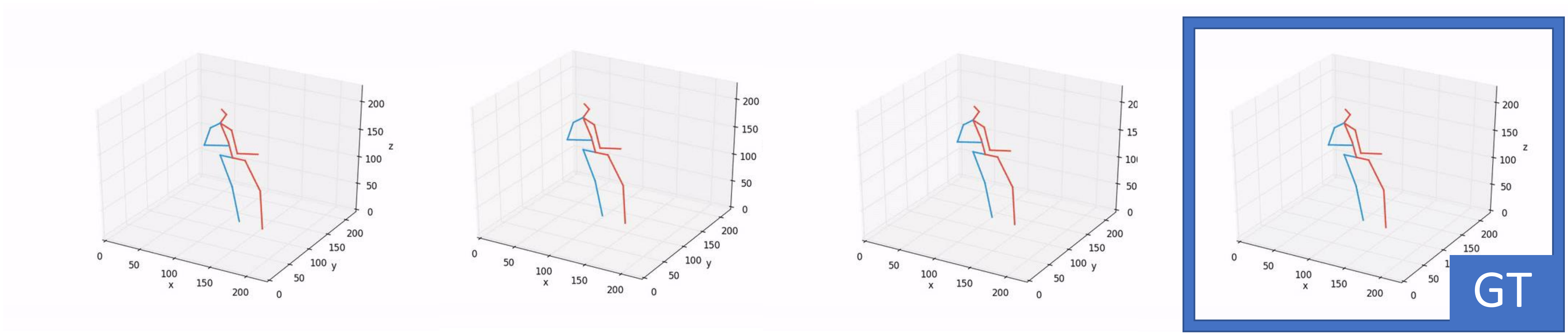
- Qualitative results



Observation: 16 frames
Generation: 60 frames

Experiments and Results

- Qualitative results



Observation: 16 frames

Generation: 160 frames

UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Tong Zhang, Nick Barnes, CVPR 2020

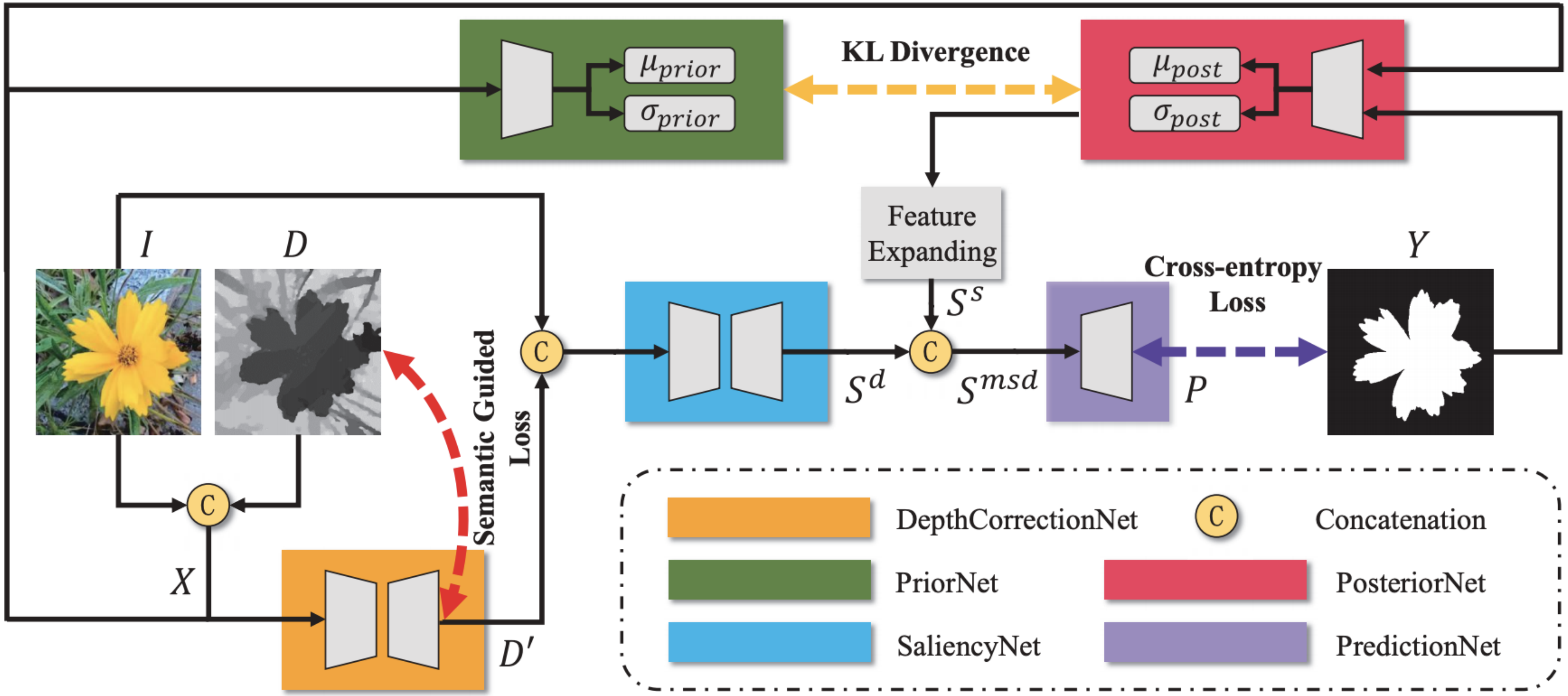
RGB-D Saliency Detection

- Task
  - Separating the most conspicuous objects that attract humans from the background
- Existing approaches
  - Treat saliency detection as a point estimation problem
  - Produce a single saliency map for each input image following a deterministic pipeline
  - Fails to capture the stochastic characteristic of saliency detection

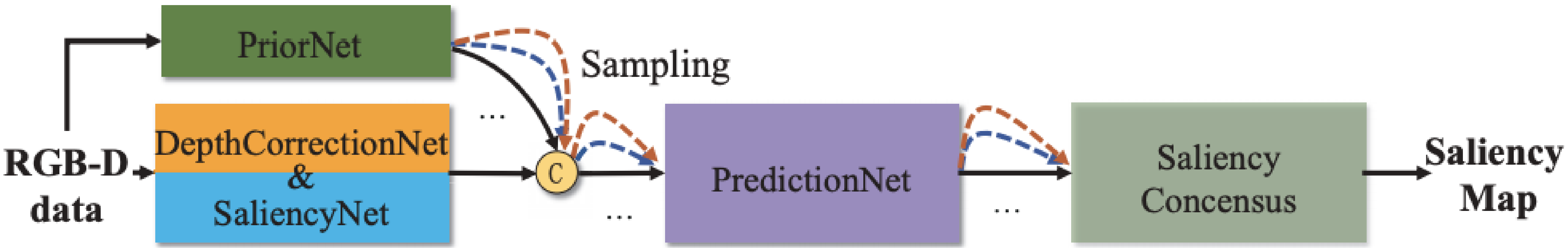
RGB-D Saliency Detection

- Our Goal
  - Employ uncertainty for RGB-D saliency detection by learning from the data labeling process
  - Interested in how the network produces multiple predictions

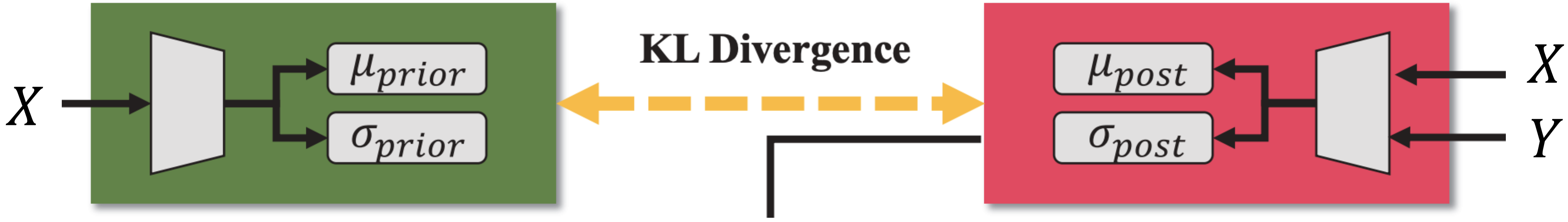
RGB-D Saliency Detection



RGB-D Saliency Detection

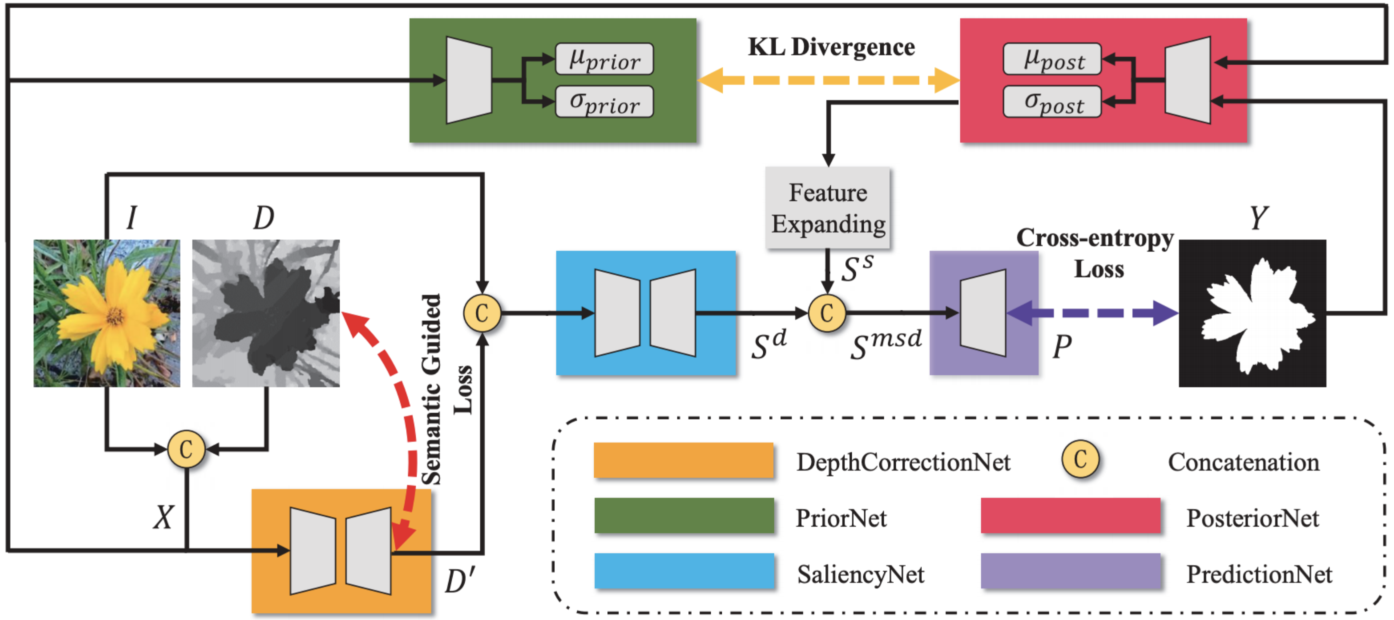


RGB-D Saliency Detection



X : Conditioning variable (RGB-D) image pair
$z \sim \mathcal{N}\left(\mu, diag(\sigma^2)\right)$: Latent variable
Y : Output variable
$P_\theta(z|X)$: Prior Net that maps the input RGB-D (X) to latent feature space
$Q_\varphi(z|X,Y)$: Posterior Net
$D_{KL}(Q_\varphi(z|X,Y)||P_\theta(z|X))$: Regularization loss to reduce the gap between the prior and posterior

RGB-D Saliency Detection



$P_\omega(Y|X,z)$: Likelihood of $P(Y)$ given latent variable z and conditioning variable X

$$\mathcal{L}_{CVAE} = E_{z\sim Q_\varphi(z|X,Y)}[-\log P_\omega(Y|X,z)] + D_{KL}(Q_\varphi(z|X,Y)||P_\theta(z|X))$$

RGB-D Saliency Detection

	Metric	DF [43]	AFNet [54]	CTMF [24]	MMCI [7]	PCF [5]	TANet [6]	CPFP [64]	DMRA [61]	*UC-Net* Ours
		Deep Models								
NJU2K [28]	$S_\alpha \uparrow$	.763	.822	.849	.858	.877	.879	.878	.886	**.897**
	$F_\beta \uparrow$	.653	.827	.779	.793	.840	.841	.850	.873	**.886**
	$E_\xi \uparrow$	.700	.867	.846	.851	.895	.895	.910	.920	**.930**
	$\mathcal{M} \downarrow$	.140	.077	.085	.079	.059	.061	.053	.051	**.043**
SSB [40]	$S_\alpha \uparrow$	.757	.825	.848	.873	.875	.871	.879	.835	**.903**
	$F_\beta \uparrow$	.617	.806	.758	.813	.818	.828	.841	.837	**.884**
	$E_\xi \uparrow$	.692	.872	.841	.873	.887	.893	.911	.879	**.938**
	$\mathcal{M} \downarrow$	.141	.075	.086	.068	.064	.060	.051	.066	**.039**
DES [8]	$S_\alpha \uparrow$	.752	.770	.863	.848	.842	.858	.872	.900	**.934**
	$F_\beta \uparrow$	.604	.713	.756	.735	.765	.790	.824	.873	**.919**
	$E_\xi \uparrow$	.684	.809	.826	.825	.838	.863	.888	.933	**.967**
	$\mathcal{M} \downarrow$	.093	.068	.055	.065	.049	.046	.038	.030	**.019**
NLPR [41]	$S_\alpha \uparrow$	.806	.799	.860	.856	.874	.886	.888	.899	**.920**
	$F_\beta \uparrow$	.664	.755	.740	.737	.802	.819	.840	.865	**.891**
	$E_\xi \uparrow$	.757	.851	.840	.841	.887	.902	.918	.940	**.951**
	$\mathcal{M} \downarrow$	.079	.058	.056	.059	.044	.041	.036	.031	**.025**
LFSD [35]	$S_\alpha \uparrow$	.791	.738	.796	.787	.794	.801	.828	.847	**.864**
	$F_\beta \uparrow$	.679	.736	.756	.722	.761	.771	.811	.845	**.855**
	$E_\xi \uparrow$	.725	.796	.810	.775	.818	.821	.863	.893	**.901**
	$\mathcal{M} \downarrow$	.138	.134	.119	.132	.112	.111	.088	.075	**.066**
SIP [18]	$S_\alpha \uparrow$	.653	.720	.716	.833	.842	.835	.850	.806	**.875**
	$F_\beta \uparrow$	.465	.702	.608	.771	.814	.803	.821	.811	**.867**
	$E_\xi \uparrow$	.565	.793	.704	.845	.878	.870	.893	.844	**.914**
	$\mathcal{M} \downarrow$	.185	.118	.139	.086	.071	.075	.064	.085	**.051**

RGB-D Saliency Detection

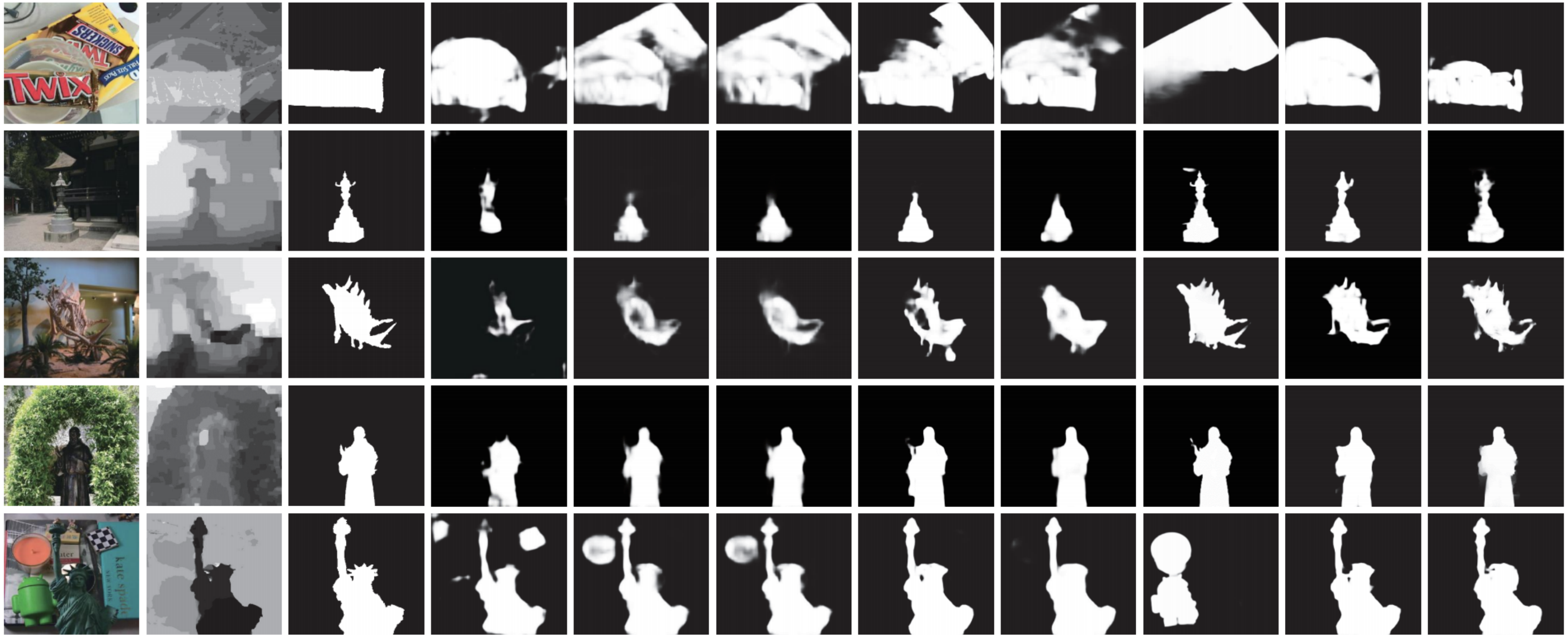


Image Depth GT DMRA MH1 MH2 DP1 DP2 Ours (1) Ours (2) UC-Net

Conclusion

- Generative models
  - Latent Variable models
  - Variational Autoencoders
- Stochastic problems
  - Human Motion Prediction
  - RGB-D Saliency Object Detection
- What we care about is:
  - High quality solutions
  - Diverse solutions where there is inherent uncertainty

Thanks!
Q&A