# Deep Latent-Variable Generative Models for Multimedia Processing

**Xiaoyu LIN**
**June 25th, 2024**

**Under the supervision of**
**Dr. Xavier Alameda-Pineda and Prof. Laurent Girin**

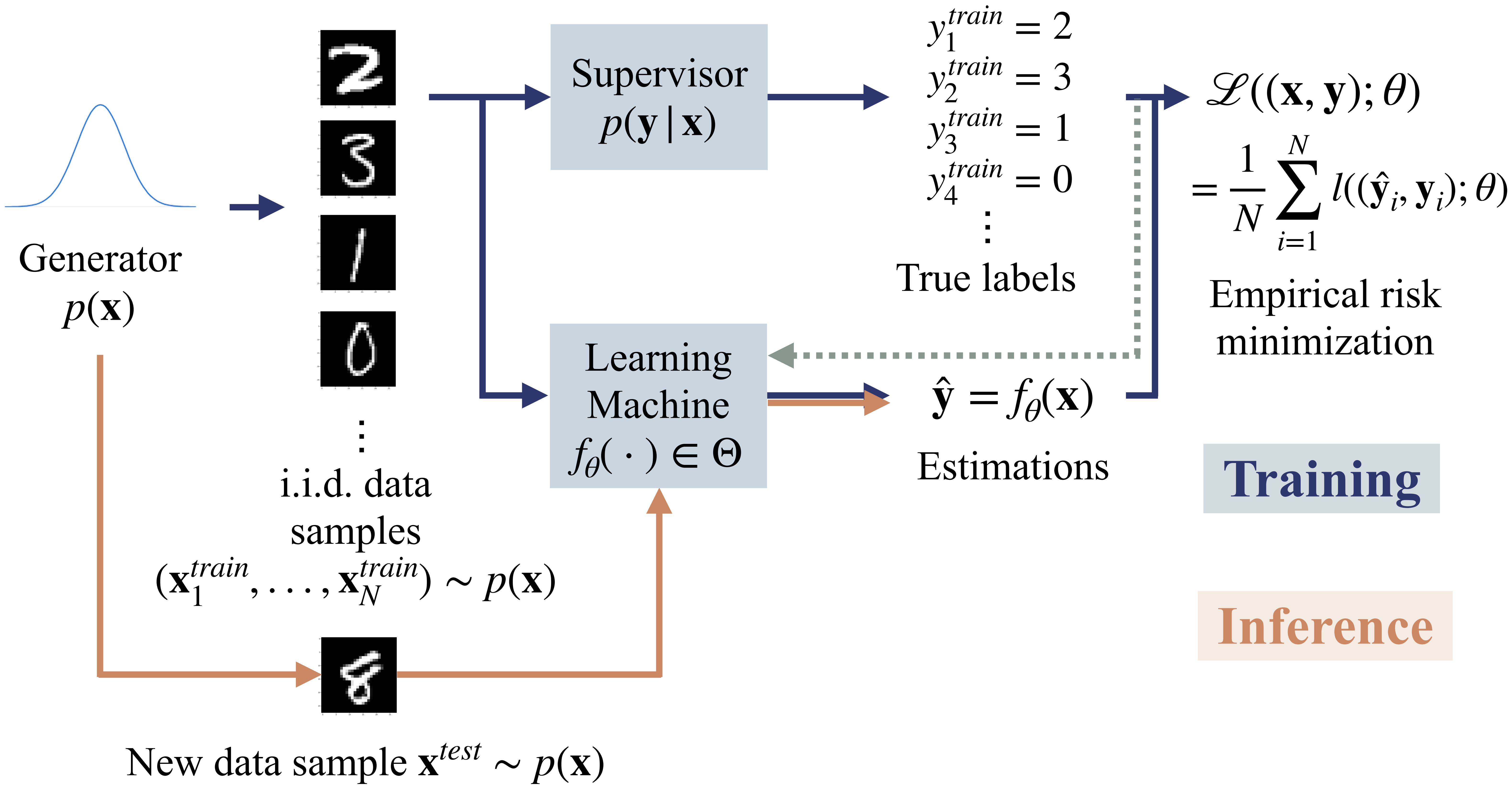
**Jury:**
**Prof. Gaël RICHARD (Rapporteur)**
**Dr. David PICARD (Rapporteur)**
**Prof. Shai BEN-DAVID (Examinateur)**
**Prof. Dorothea KOLOSSA (Examinatrice)**
**Prof. Jean-Marc BROSSIER (Examinateur)**

# CONTENTS

# 01. Introduction

# What makes the great success of today's AI systems?

## Statistical learning framework[1]




### Key factors of success[2]

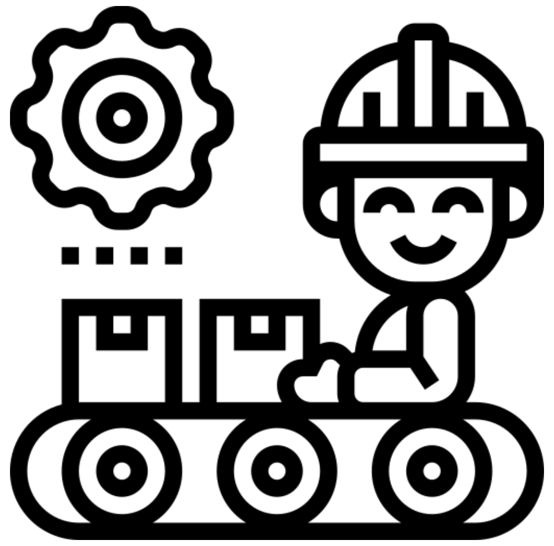
- Large dataset
- Well-designed learning machine
- Computational ability
- The i.i.d. data assumption

$$(\mathbf{x}^{train}, \mathbf{y}^{train}) \sim p(\mathbf{x}, \mathbf{y})$$
$$(\mathbf{x}^{test}, \mathbf{y}^{test}) \sim p(\mathbf{x}, \mathbf{y})$$

[1] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. 2000.
[2] Bernhard Schölkopf, and Julius von Kügelgen. From statistical to causal learning. *Proc. of the Int.Congress of Mathematicians.* 2022.

# In what situations does this system not work?

1. When we do not have enough data for training


**Health care** **Industrial production** **Finance**

**ImageNet[3]**
Object recognition
~1,200,000 images

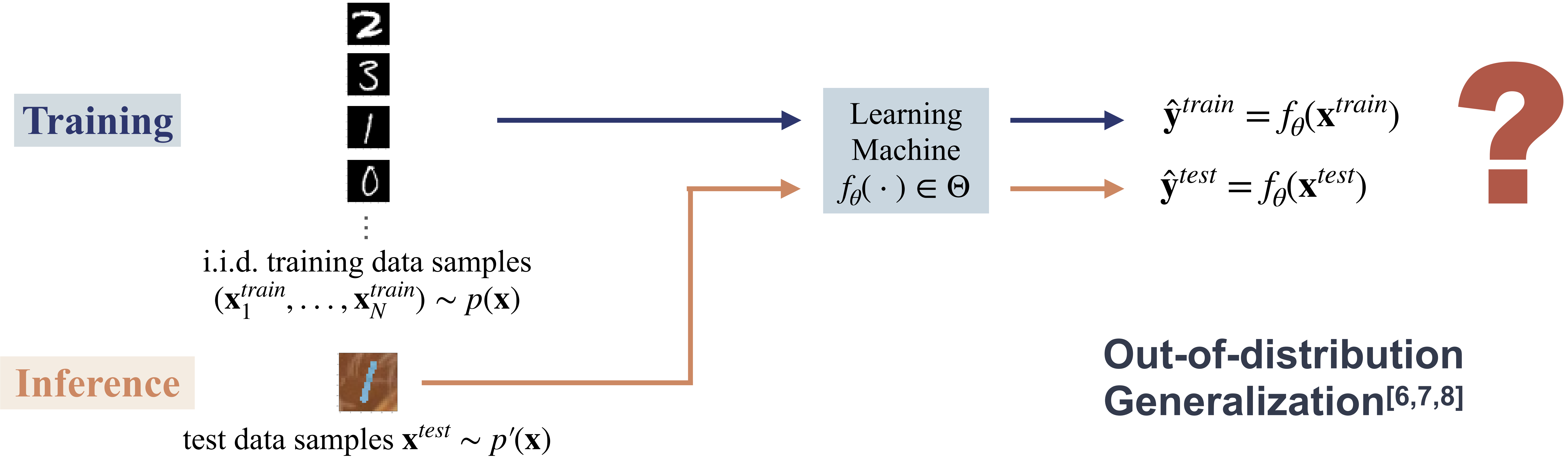
**GPT3[4]**
Text generation
~570 GB pf text data


**Whisper[5]**
Speech recognition
~680,000 hours of audio

[3] Jia Deng, et al. ImageNet: A large-scale hierarchical image database. *Proc. IEEE Int. Conf. Computer Vision Pattern Recogn. (CVPR).* 2009.
[4] Tom B. Brown, et al. Language models are few-shot learners. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2020.
[5] Alec Radford, et al. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356.* 2022.

# In what situations does this system not work?

1. When we do not have enough data for training
2. When during inference, the new data $(\mathbf{x}^{test}, \mathbf{y}^{test})$ does not follow distribution $p(\mathbf{x}, \mathbf{y})$

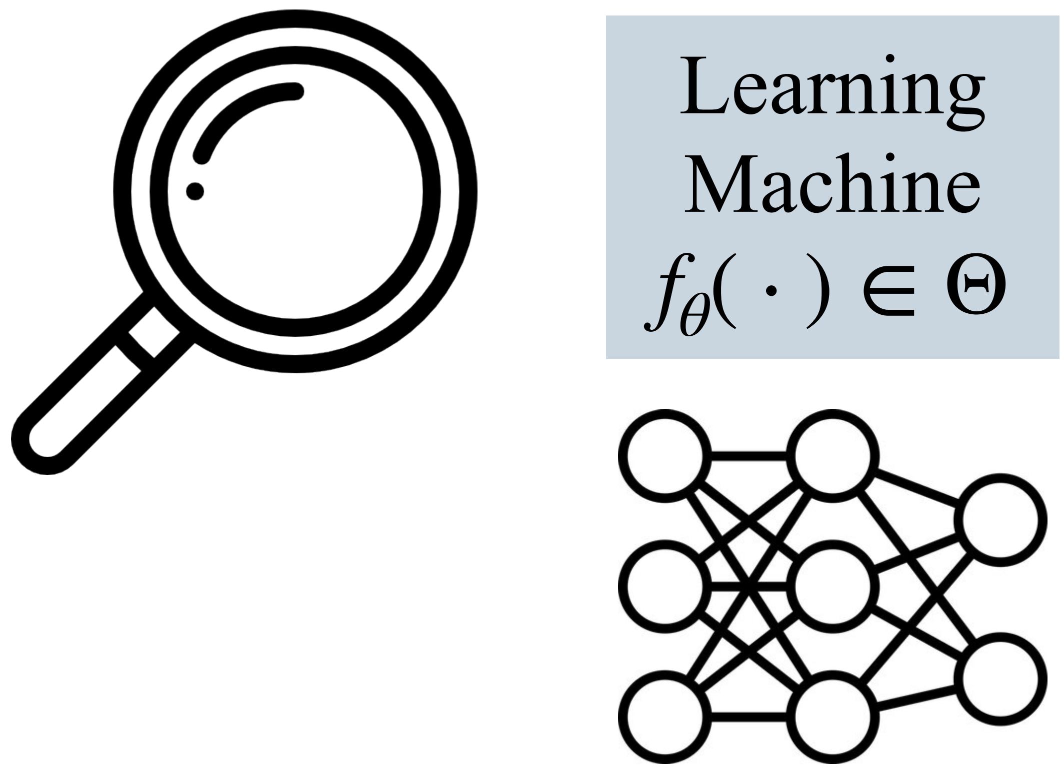



**Out-of-distribution Generalization[6,7,8]**

[6] Shai Ben-David, et al. A theory of learning from different domains. *Mach. Learn.* 2010.
[7] Krikamol Muandet, et al. Domain Generalization via Invariant Feature Representation. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2013.
[8] Jiashuo Liu, et al. Towards Out-Of-Distribution Generalization: A Survey. *arXiv preprint arXiv:2108.13624.* 2021.

# In what situations does this system not work?

1. When we do not have enough data for training
2. When during inference, the new data $(\mathbf{x}^{test}, \mathbf{y}^{test})$ does not follow distribution $p(\mathbf{x}, \mathbf{y})$
3. When we would like to understand the "black-box" learning machine $f_\theta(\cdot)$




=

**Interpretable AI[9,10]**

[9] Been Kim, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proc. Int. Conf. Mach. Learn. (ICML).* 2018.
[10] Finale Doshi-Velez, et al. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608.* 2017.

# Background of the proposed solution

## Statistical learning framework (ERM inductive principle)

Learning Machine $f_\theta(\cdot) \in \Theta$ $\approx$ Supervisor $p(\mathbf{y} \mid \mathbf{x})$ $\Rightarrow$ $\hat{\mathbf{y}} = f_\theta(\mathbf{x}) \approx \mathbb{E}[\mathbf{y} \mid \mathbf{x}]$

Empirical risk minimization

## Bayesian inference

likelihood

prior

$$\underbrace{p_\theta(\mathbf{y} \mid \mathbf{x})}_{\text{posterior}} = \frac{p_\theta(\mathbf{x} \mid \mathbf{y}) p_\theta(\mathbf{y})}{\int p_\theta(\mathbf{x} \mid \mathbf{y}) p_\theta(\mathbf{y}) dy}$$

marginal likelihood / evidence

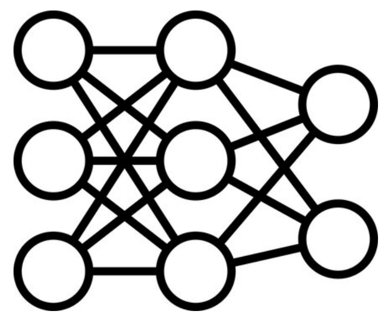
# Background of the proposed solution

## Bayesian inference

likelihood

prior

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{x} \mid \mathbf{y}) p_\theta(\mathbf{y})}{\int p_\theta(\mathbf{x} \mid \mathbf{y}) p_\theta(\mathbf{y}) d\mathbf{y}}$$

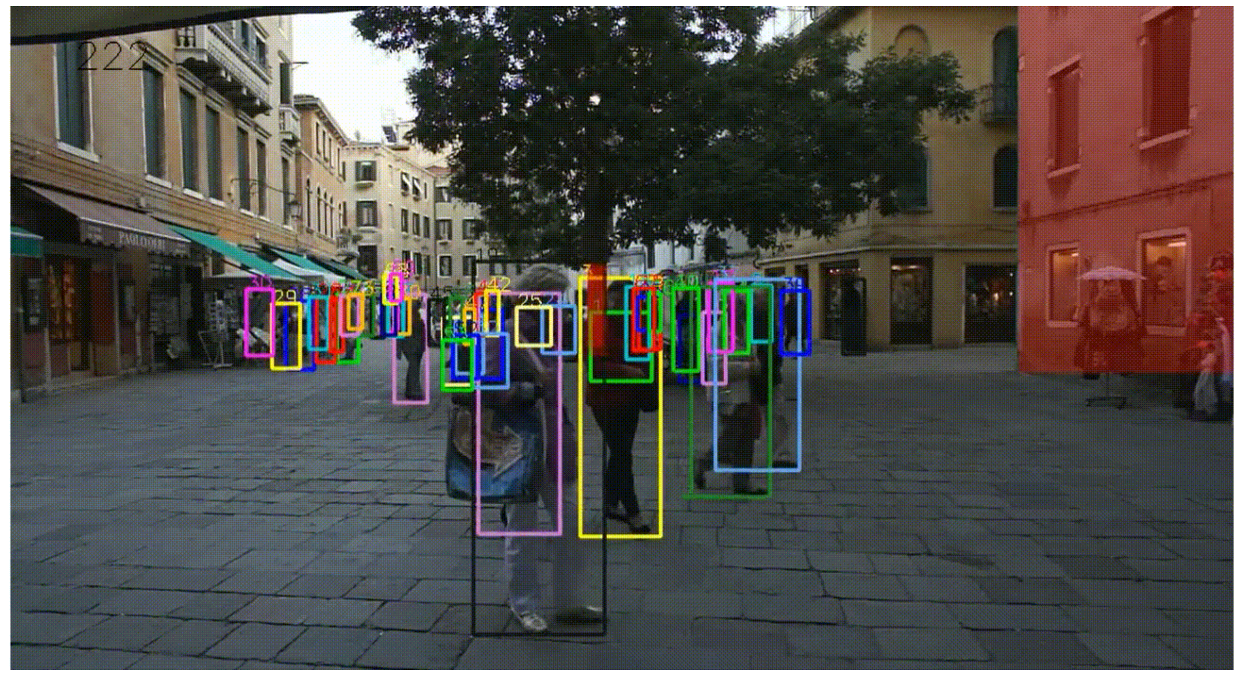


posterior

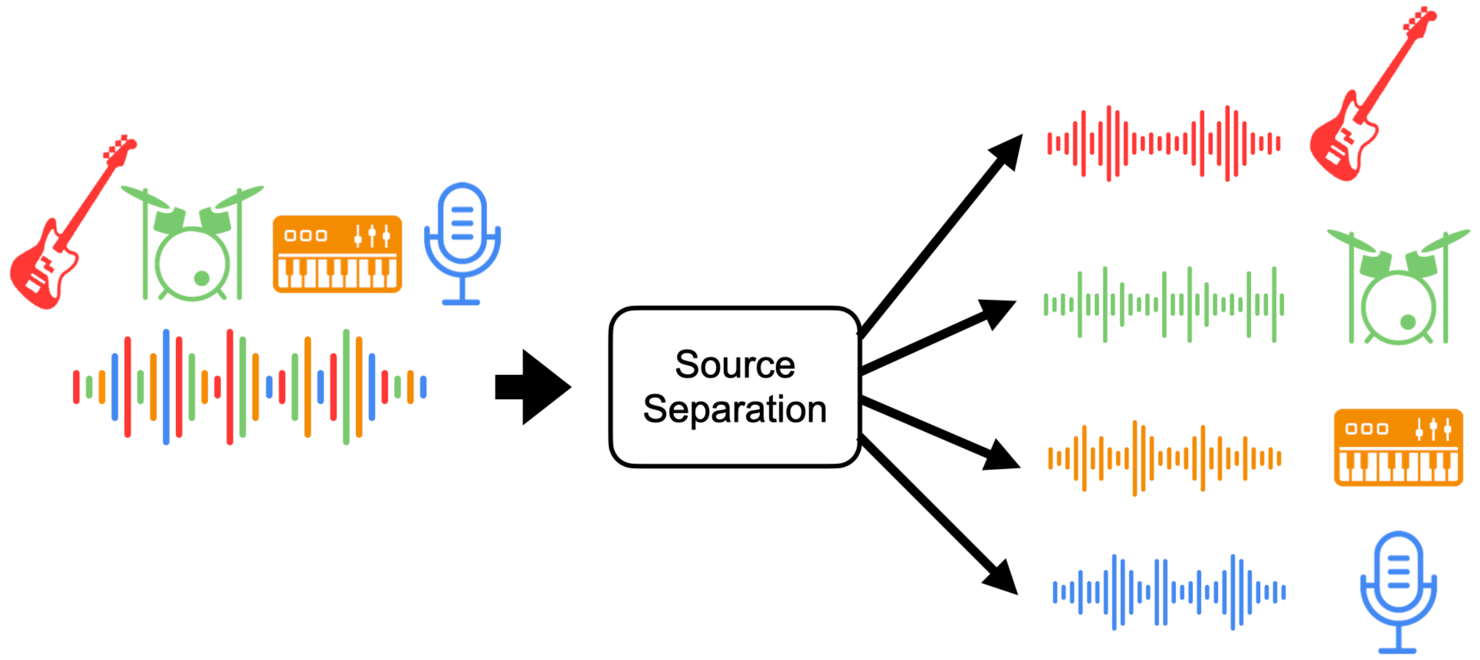
marginal likelihood / evidence

- Model $p_\theta(\mathbf{x} \mid \mathbf{y})$ with domain specific knowledge.
- Model $p_\theta(\mathbf{y})$ with a deep probabilistic generative model.
- Infer $p_\theta(\mathbf{y} \mid \mathbf{x})$ with Bayesian inference methodology.

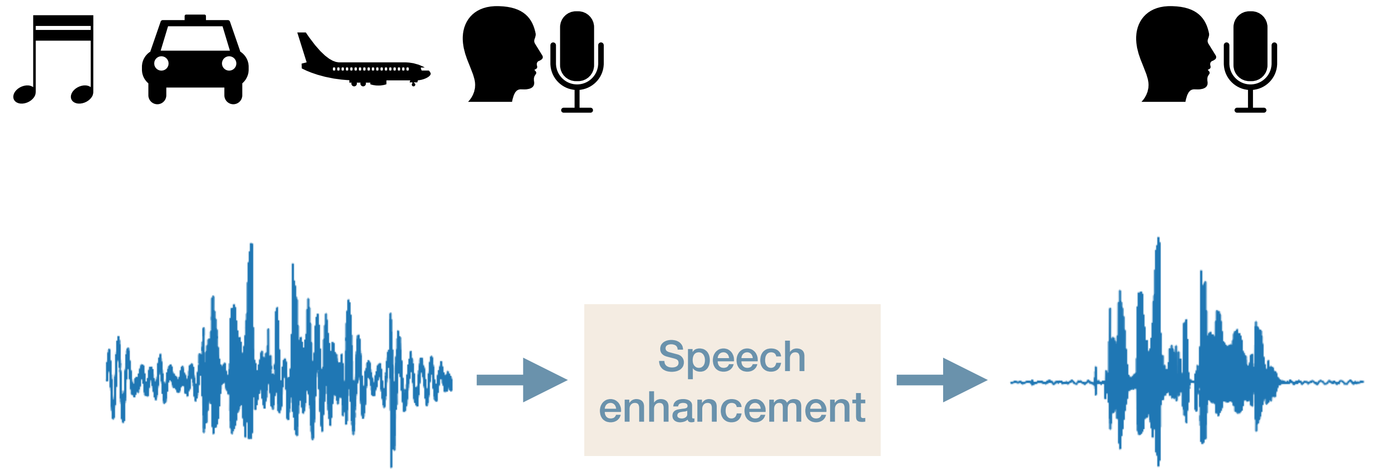
# Application to three multimedia processing tasks



**Multi-Object Tracking**




**Single-Channel Audio Source Separation**

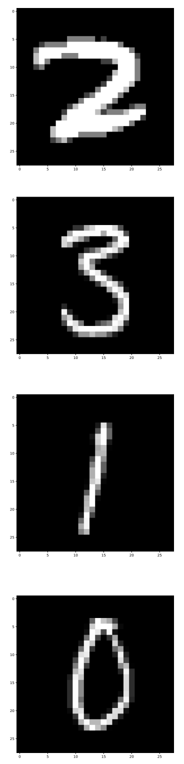



**Speech Enhancement**

# 02.
# Methodological Background

# What are probabilistic generative models?

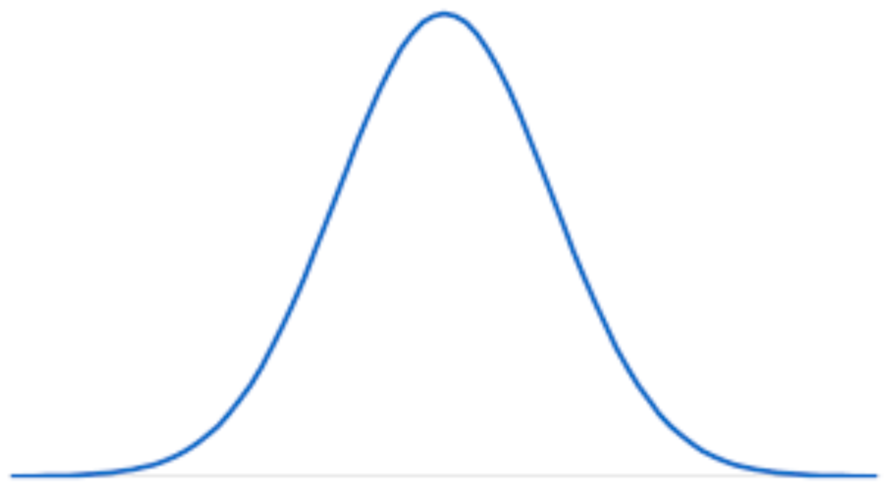
The probabilistic generative models aim to estimate the probability distribution $p(\mathbf{x})$, given a set of i.i.d. data samples $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.



$\vdots$

i.i.d. data samples
$(\mathbf{x}_1, \ldots, \mathbf{x}_N)$

$\sim$

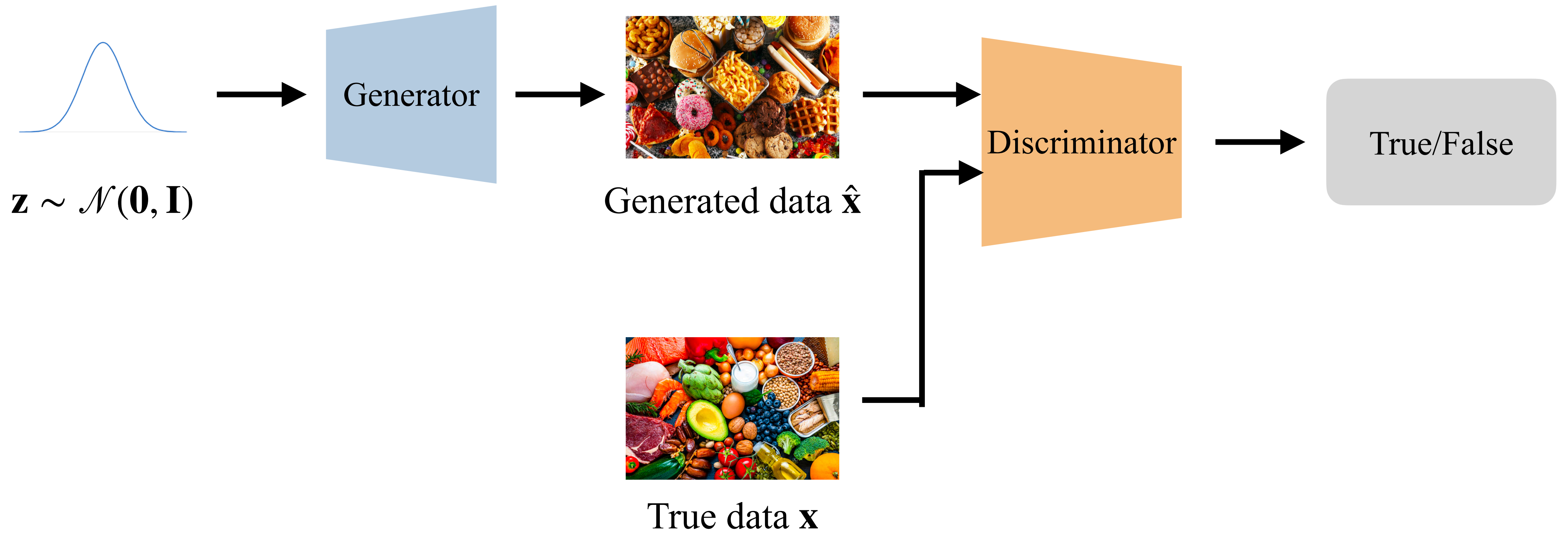


Generator
$p(\mathbf{x})$

# Different types of probabilistic generative models

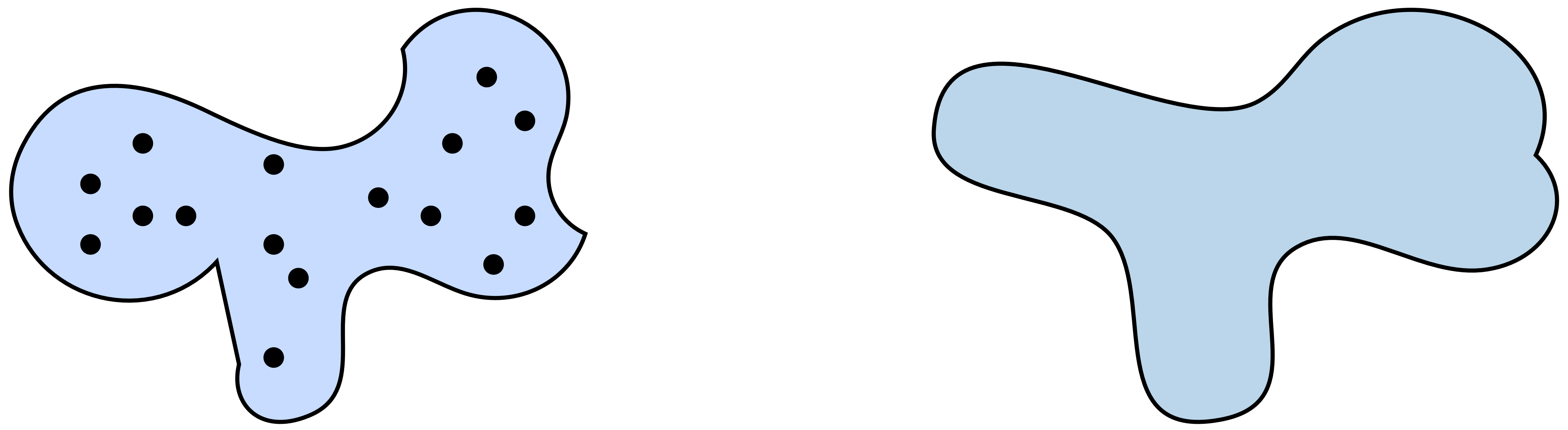
## Implicit generative models

Generative Adversarial Networks[11]




## Explicit generative models

Explicitly model the probability density function[12, 13, 14, 15]



True data distribution $p_{data}(\mathbf{x})$

Parametric probabilistic model $p_\theta(\mathbf{x})$

Maximize log-likelihood: $\mathcal{L}(\mathbf{x}; \theta) = \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x_i})$

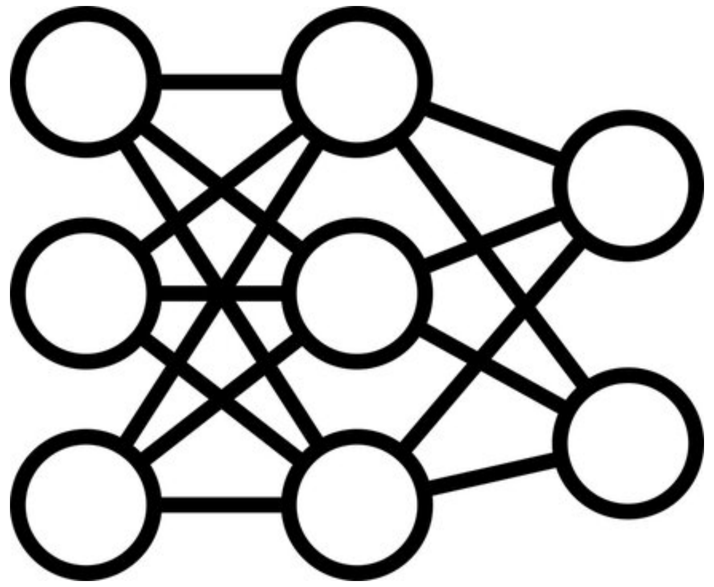
[11] Ian Goodfellow, et al. Generative adversarial nets. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2014.
[12] Benigno Uria, et al. Neural autoregressive distribution estimation. *J. Mach. Learn. Res.* 2016.
[13] Diederik P. Kingma, et al. Improved variational inference with inverse autoregressive flow. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2016.
[14] Yee Whye Teh, et al. Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.* 2003.
[15] Jonathan Ho, et al. Denoising diffusion probabilistic models. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2020.

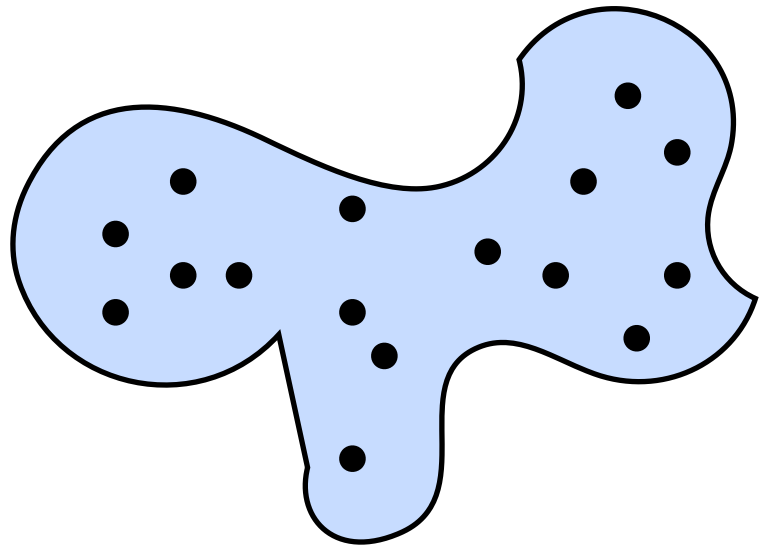
# Different types of probabilistic generative models

## Explicit generative models

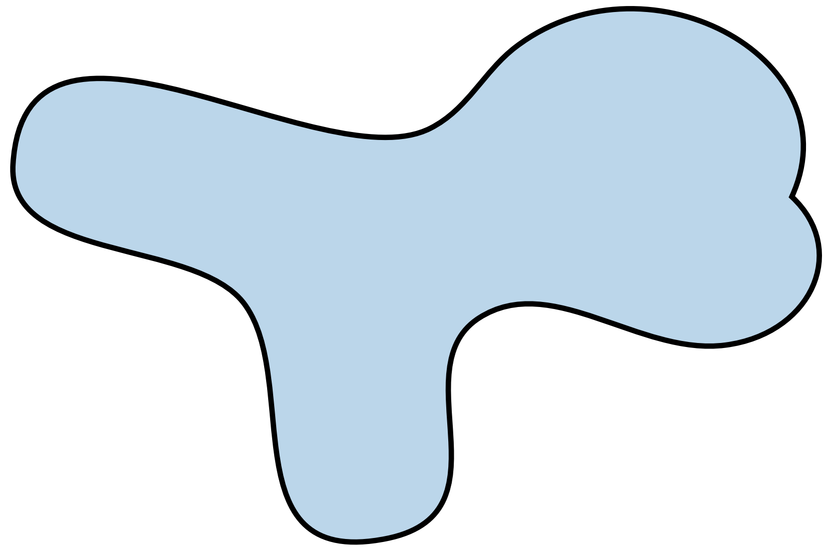
Explicitly model the probability density function[12, 13, 14, 15]

$p_\theta(\mathbf{x})$





True data distribution
$p_{data}(\mathbf{x})$



Parametric probabilistic model
$p_\theta(\mathbf{x})$

Maximize log-likelihood: $\mathscr{L}(\mathbf{x};\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(\mathbf{x_i})$

**Deep probabilistic generative models**

Diffusion models[15]

Deep auto-regressive models[12]

Deep energy-based models[14]

Normalizing flows[13]

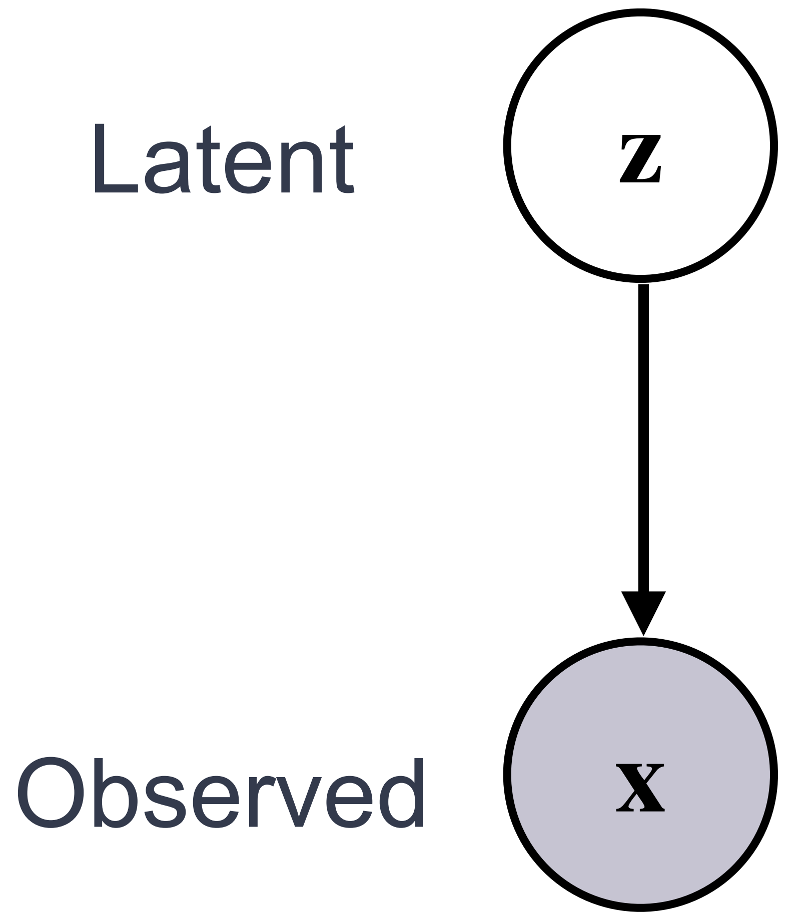
Deep latent variable models[16]

[12] Benigno Uria, et al. Neural autoregressive distribution estimation. *J. Mach. Learn. Res.* 2016.
[13] Diederik P. Kingma, et al. Improved variational inference with inverse autoregressive flow. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2016.
[14] Yee Whye Teh, et al. Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.* 2003.
[15] Jonathan Ho, et al. Denoising diffusion probabilistic models. *Advances in Neural Inform. Process. Systems (NeurIPS).* 2020.
[16] Diederik P. Kingma, et al. Auto-encoding variational Bayes. *Proc. Int. Conf. Learn. Repres. (ICLR).* 2014.

# A specific type of explicit generative models

## Latent Variable Models (LVMs)

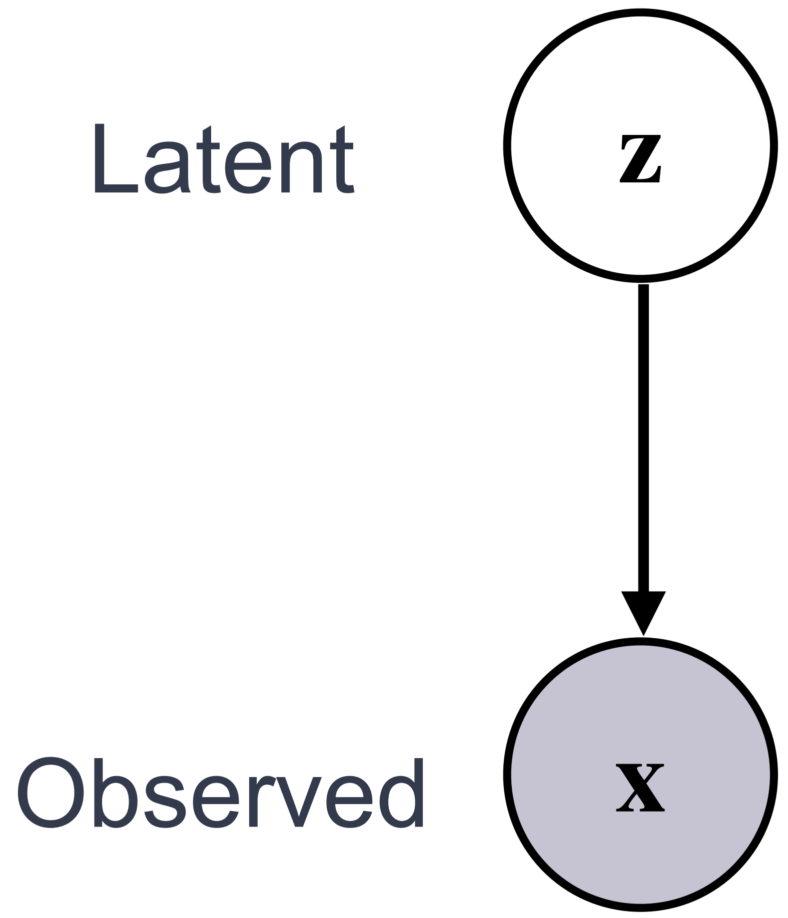
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$$

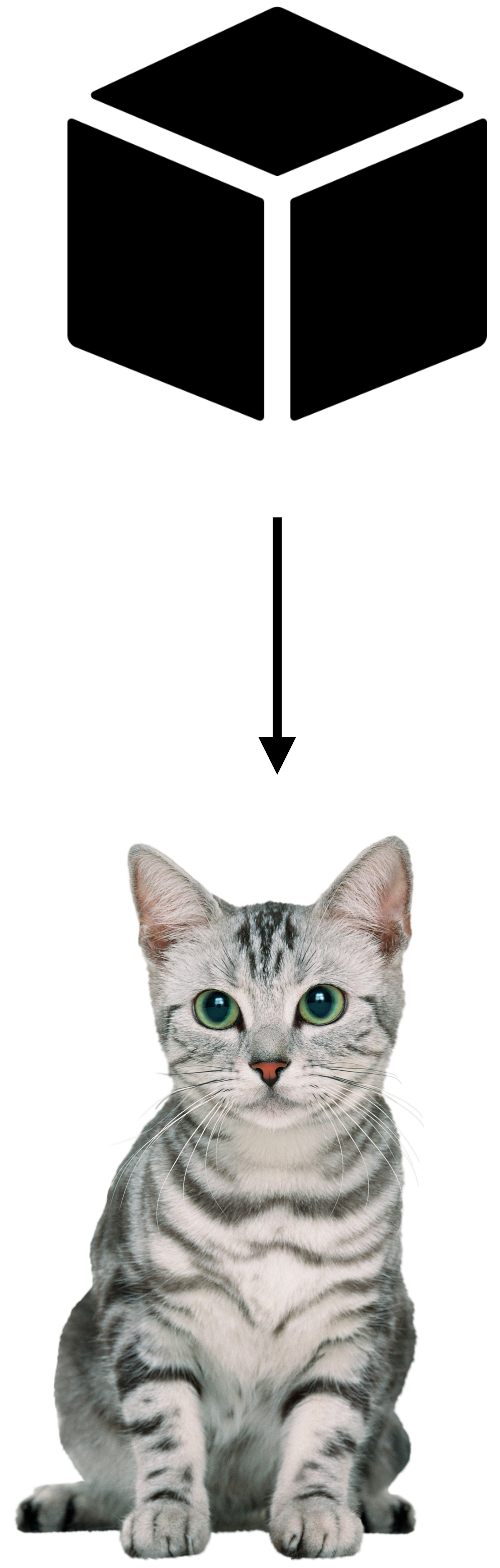


# A specific type of explicit generative models

## Latent Variable Models (LVMs)

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z}) dz$$

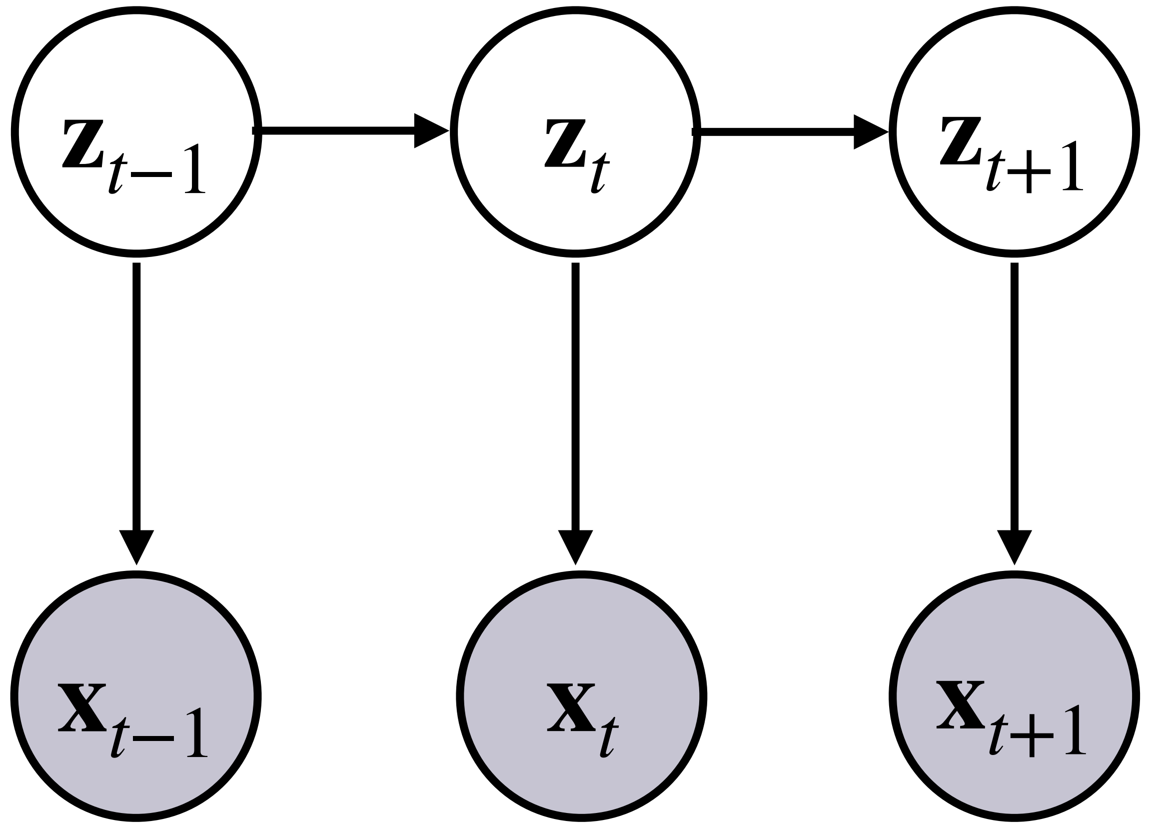


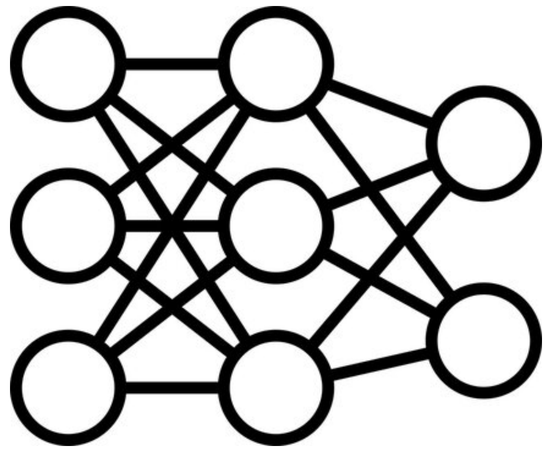
# Example: probabilistic sequential data models

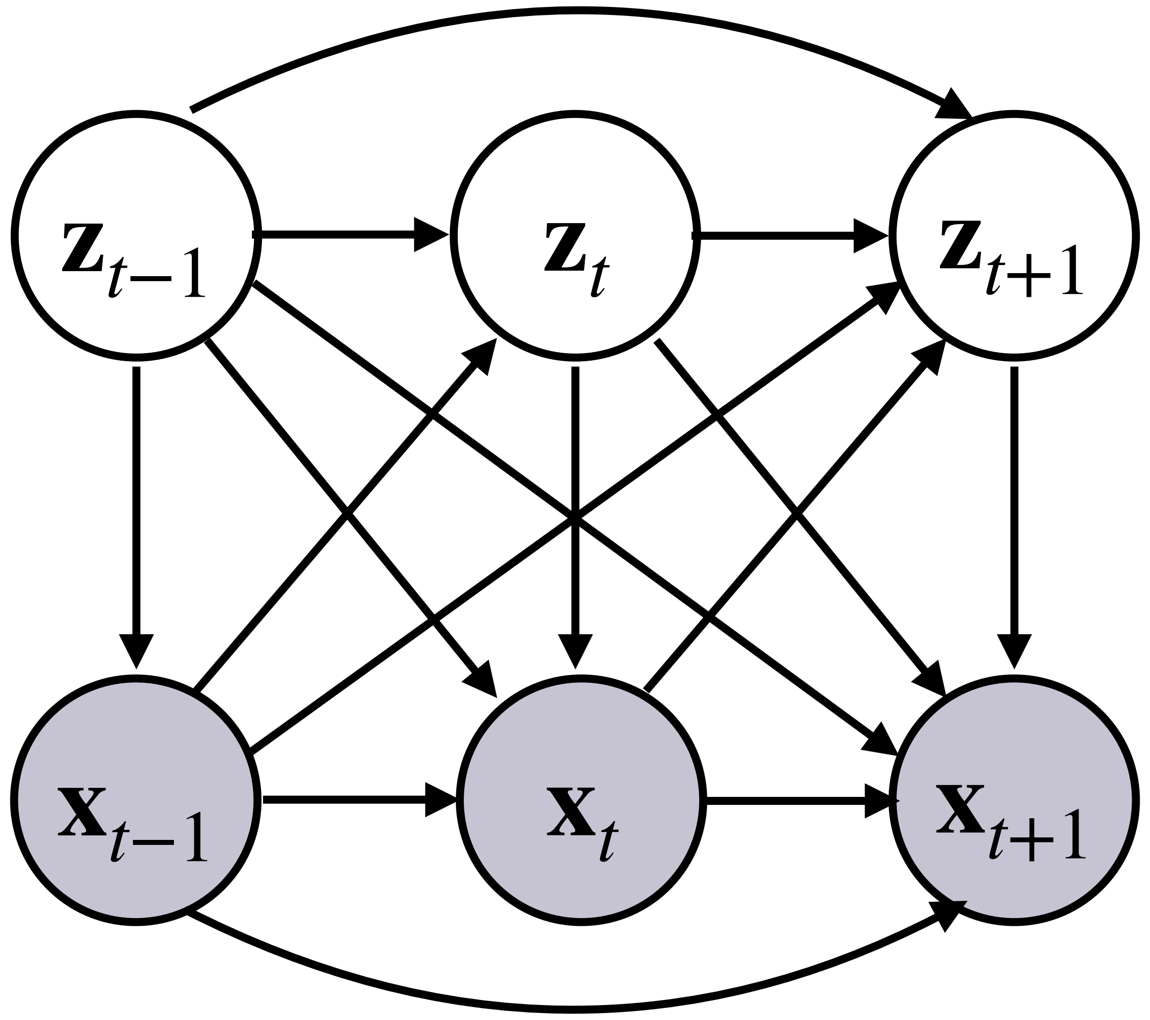
Latent sequence $\mathbf{z}_{1:T}$

Observed sequence $\mathbf{x}_{1:T}$




Non-linear dynamics






$$p_\theta(\mathbf{x}_{1:T}) = \int p_\theta(\mathbf{z}_1) \prod_{t=2}^{T} p_\theta(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_t \mid \mathbf{z}_t) d\mathbf{z}_{1:T}$$

$$p_\theta(\mathbf{x}_{1:T}) = \int p(\mathbf{x}_1, \mathbf{z}_1) \prod_{t=2}^{T} p_\theta(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}) p_\theta(\mathbf{z}_t \mid \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}) d\mathbf{z}_{1:T}$$

State Space Models (SSM)[17]

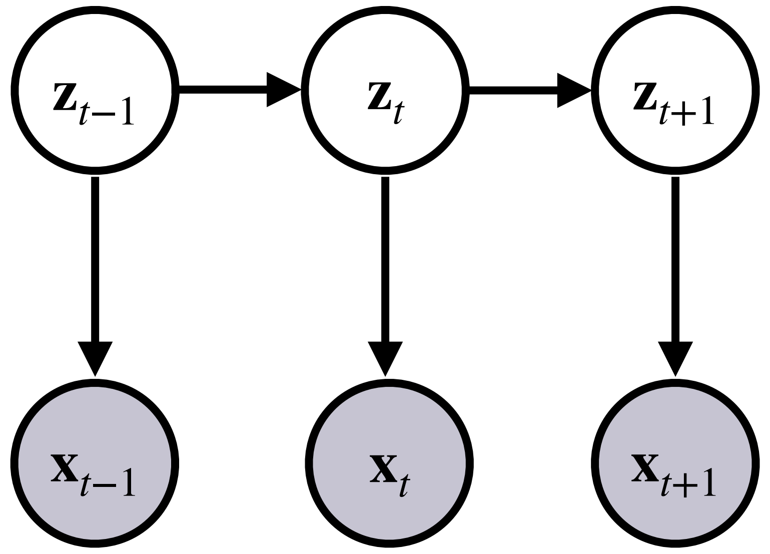
$\mathbf{z}$ discrete → Hidden Markov Model (HMM)

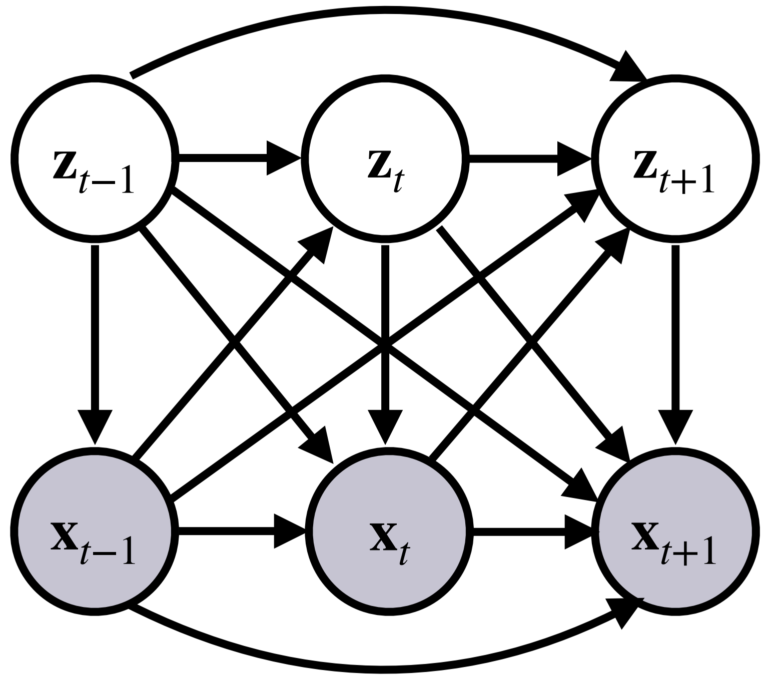
$\mathbf{z}$ continuous and linear dynamics → Linear Dynamical System (LDS)

Dynamical Variational Auto-encoders (DVAEs)[18,19,20,21]

[17] Christopher M. Bishop. Pattern Recognition and Machine Learning. 2006.
[18] Rahul Krishnan, et al. Deep kalman filters. *Advances in Approx. Bayesian Infer*. 2015.
[19] Marco Fraccaro, et al. Sequential neural models with stochastic layers. *Advances in Neural Inform. Process. Systems (NeurIPS)*. 2016.
[20] Yingzhen Li, et al. Disentangled sequential autoencoder. *Proc. Int. Conf. Mach. Learn. (ICML). 2018.*
[21] Laurent Girin, et al. Dynamical variational autoencoders: A comprehensive review. *Found. Trends Mach. Learn.* 2021.

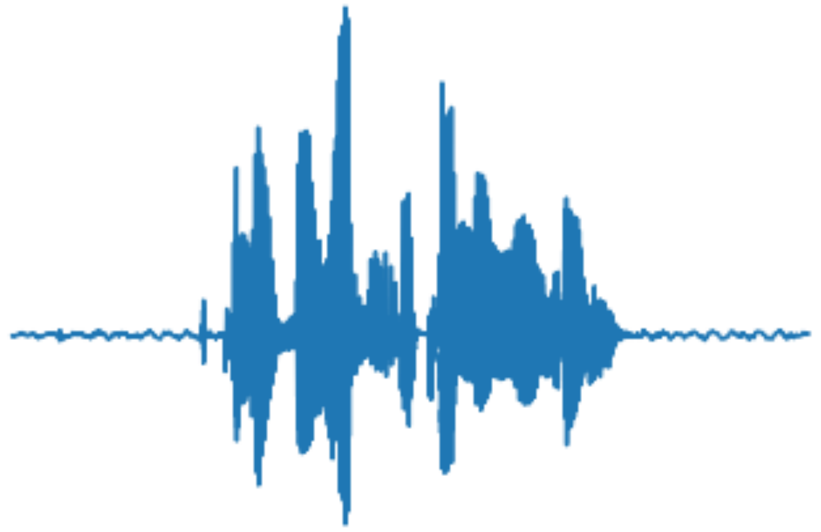
# Example: probabilistic sequential data models



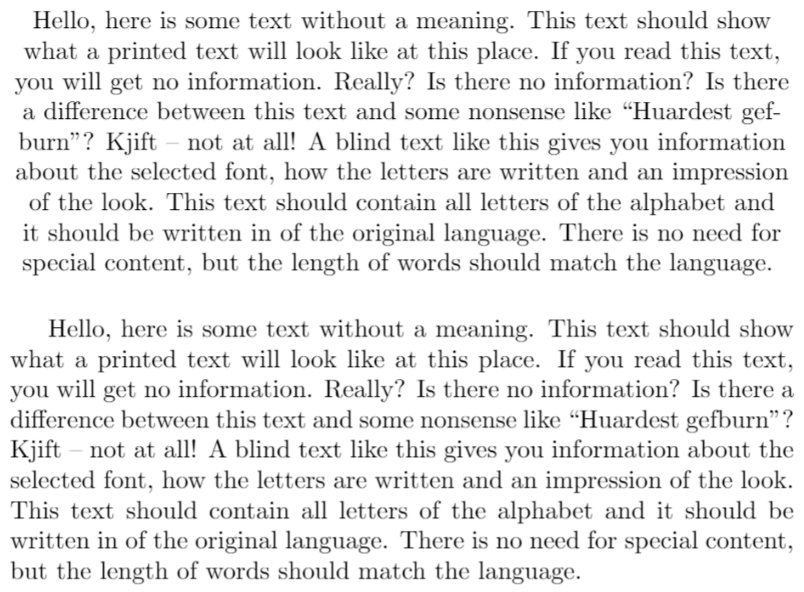




Help us to model and understand complex real-world data.

Video



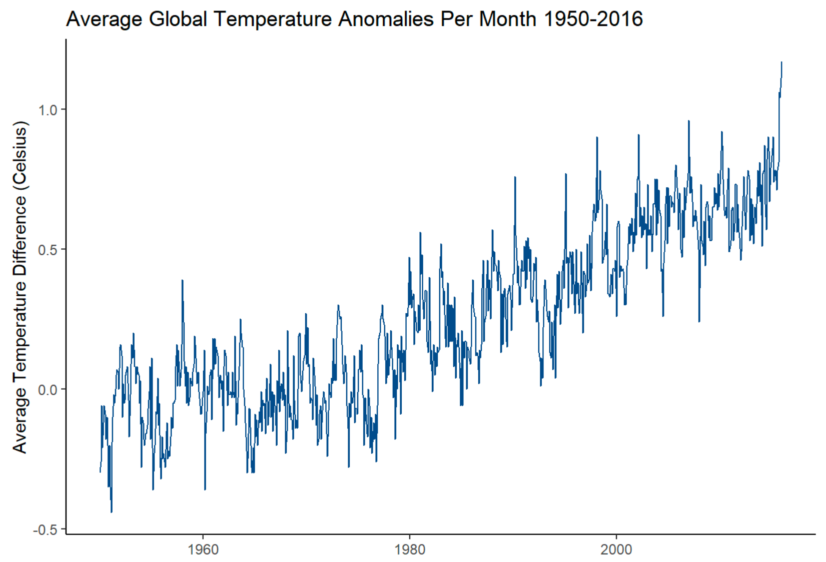
Audio



Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

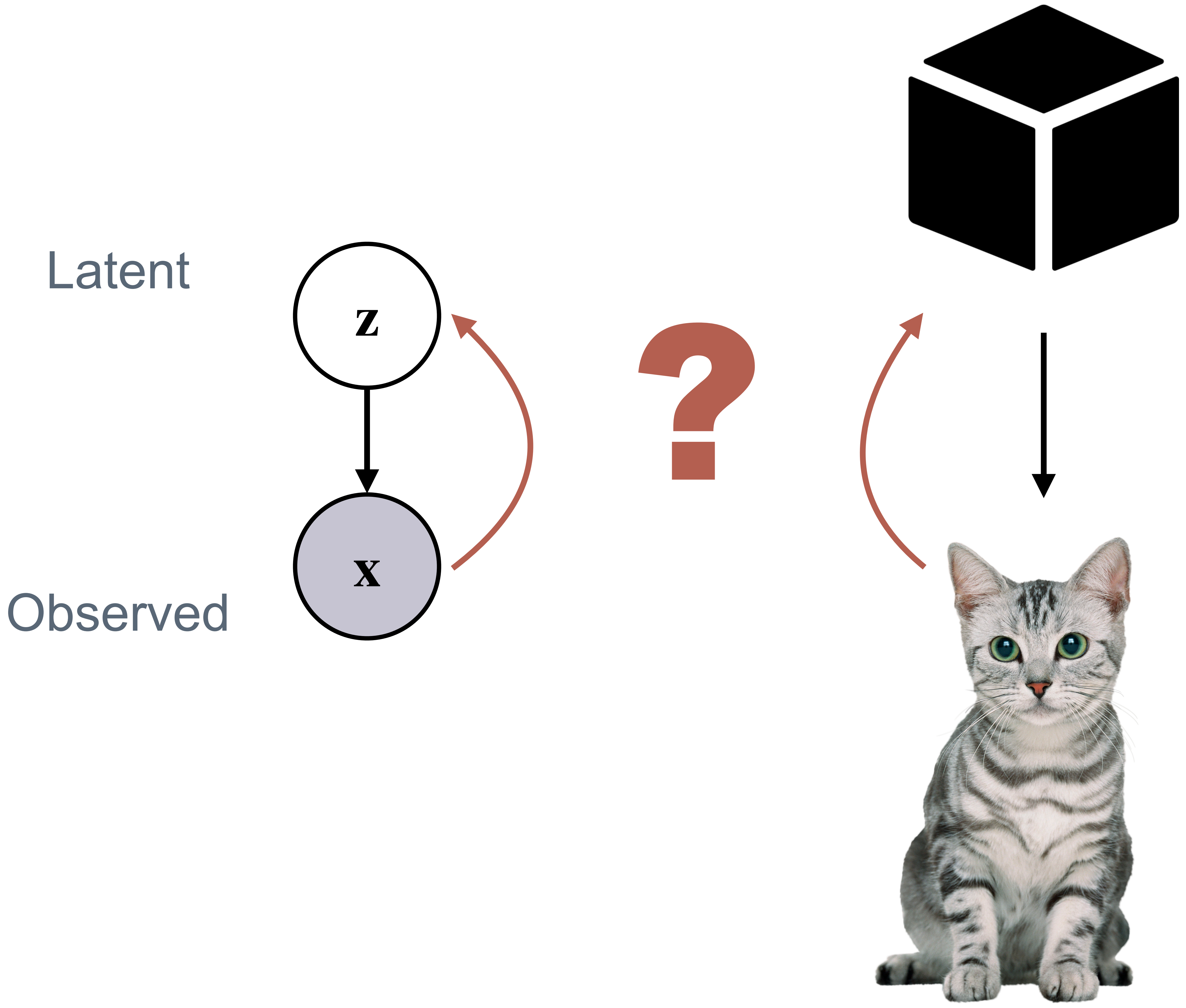
Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Text




Time series

# Another perspective of LVMs: inferring the latent variables



# Another perspective of LVMs: inferring the latent variables

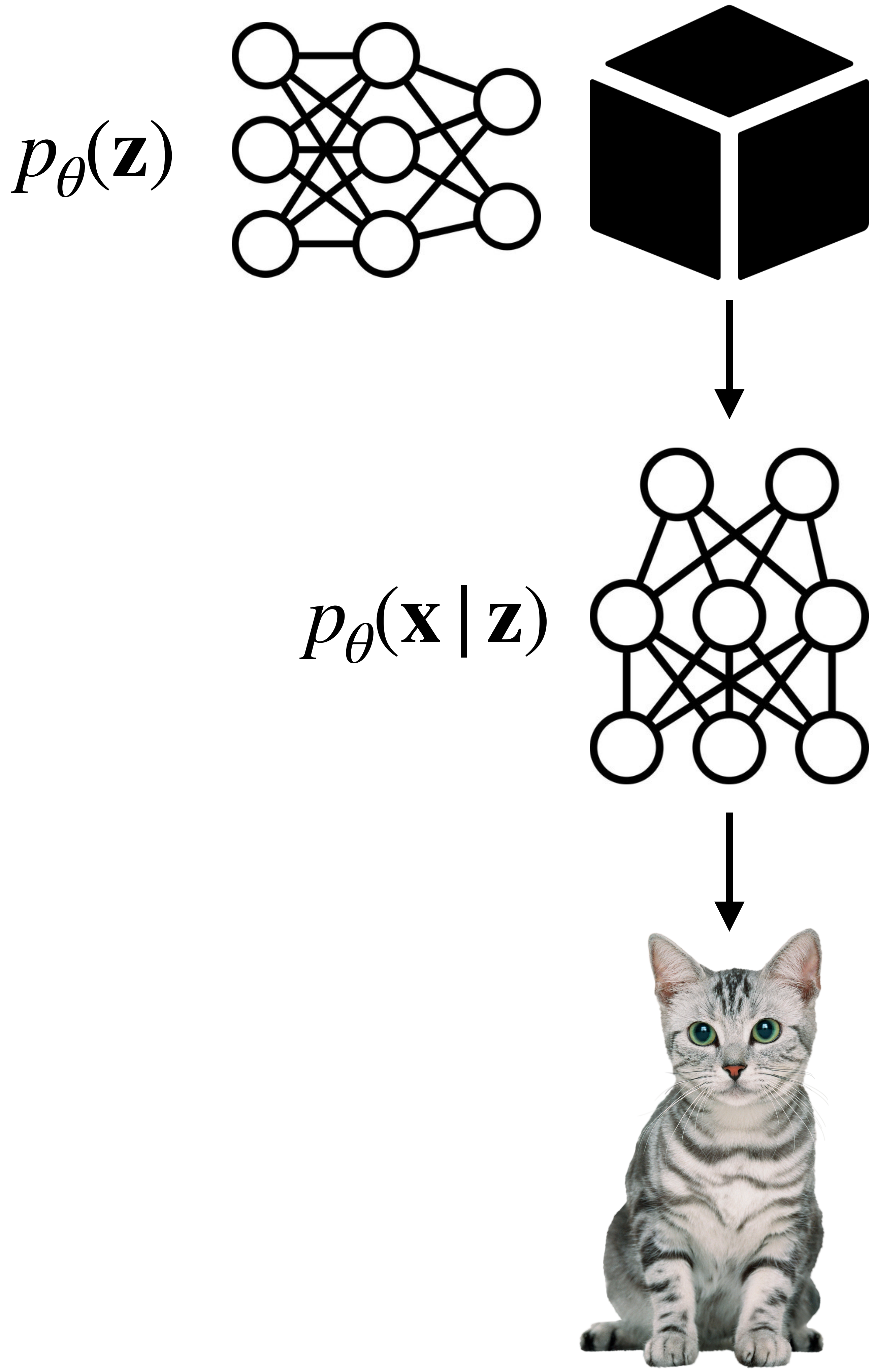
Infer the unknown latent variables: Bayesian Inference

likelihood prior

$$p_\theta(\mathbf{z} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z})}{\int p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}) dz}$$

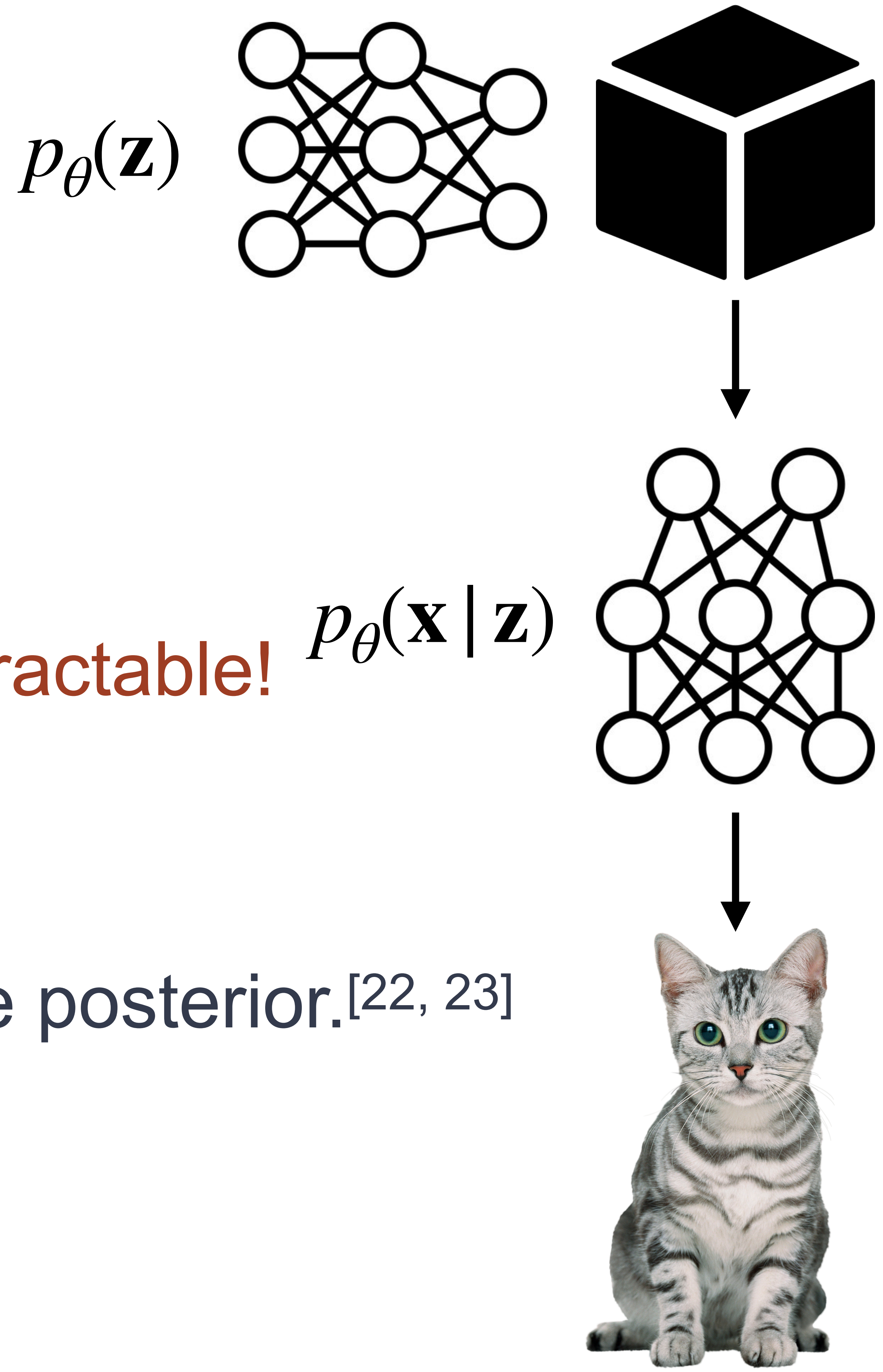
posterior

marginal likelihood / evidence



# Another perspective of LVMs: inferring the latent variables

Infer the unknown latent variables: Bayesian Inference




likelihood prior

$$p_\theta(\mathbf{z}\,|\,\mathbf{x}) = \frac{p_\theta(\mathbf{x}\,|\,\mathbf{z})p_\theta(\mathbf{z})}{\int p_\theta(\mathbf{x}\,|\,\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}}$$

posterior

Intractable!

marginal likelihood / evidence

**Solution**: introduce a variational distribution to approximate the posterior.[22, 23]

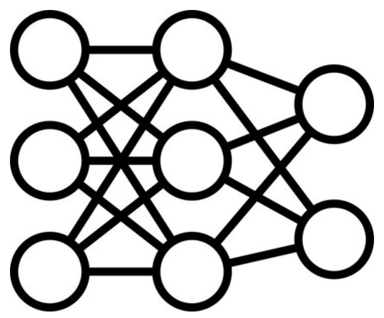
$$q(\mathbf{z}) \approx p_\theta(\mathbf{z}\,|\,\mathbf{x})$$

[22] Martin J. Wainwright, et al. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 2008.
[23] David M. Blei, et al. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 2017

# Variational inference and parameter estimation

**Maximize ELBO**: $\mathcal{L}(q,\theta) = \mathbb{E}_{q(\mathbf{z})}[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z})] \leq \log p_\theta(\mathbf{x})$

- Mean-field approximation:[24] $q(\mathbf{z}) = \prod_{i=1}^{M} q_i(z_i \,|\, \mathbf{x})$ $\longrightarrow$ Variational EM algorithm[25]
- Amortized inference:[26] $\mathcal{L}(\phi,\theta) = \mathbb{E}_{q_\phi(\mathbf{z})}[\log p_\theta(\mathbf{x}\,|\,\mathbf{z})] - KL\big(q_\phi(\mathbf{z})\,||\,p_\theta(\mathbf{z})\big)$ $\longrightarrow$ VAE[16,27]



[16] Diederik P. Kingma, et al. Auto-encoding variational Bayes. *Proc. Int. Conf. Learn. Repres. (ICLR)*. 2014.
[24] Giorgio Parisi. Statistical Field Theory. 1988.
[25] Michael I. Jordan, et al. An introduction to variational methods for graphical models. *Mach. Learn.* 1999.
[26] Samuel J. Gershman, et al. Amortized inference in probabilistic reasoning. *Proc. Annual Meeting of the Cognitive Science Society*. 2014
[27] Danilo Jimenez Rezende, et al. Stochastic backpropagation and approximate inference in deep generative models. *Proc. Int. Conf. Mach. Learn. (ICML)*. 2014.

# 03.
# Main Work

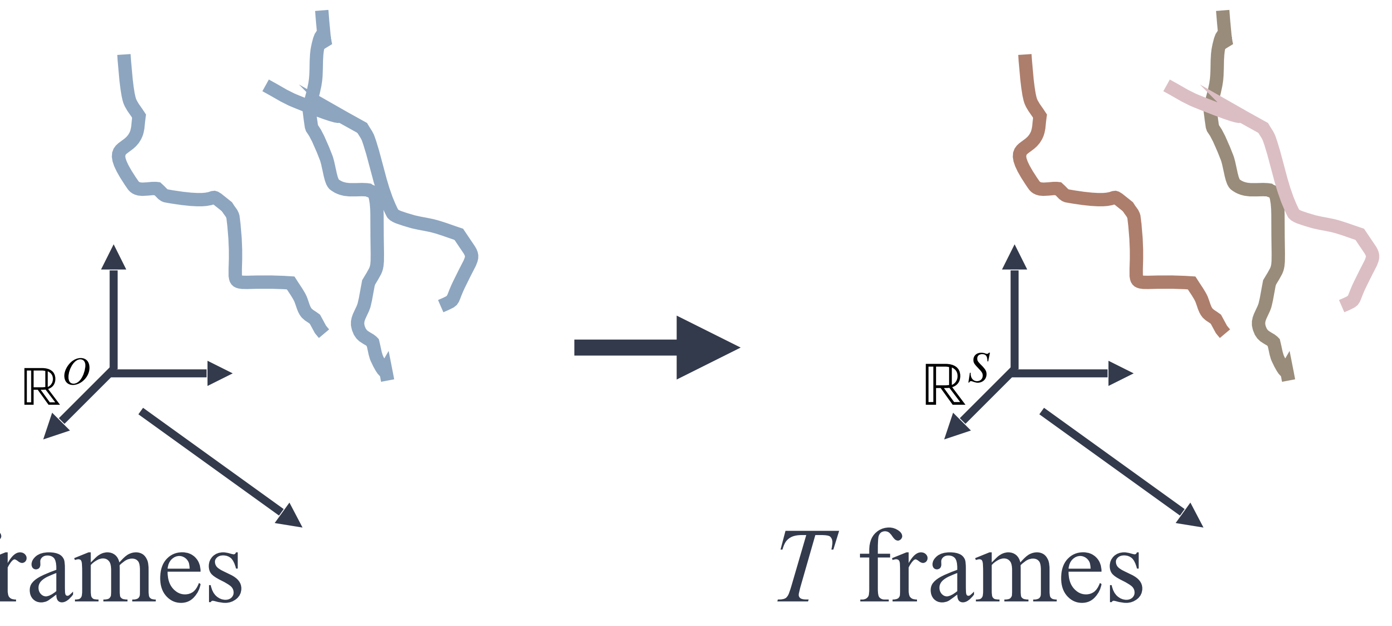
# Part 1

# Mixture of DVAEs for multi-source trajectory modeling and separation

Xiaoyu Lin, Laurent Girin, and Xavier Alameda-Pineda. “Mixture of dynamical variational autoencoders for multi-source trajectory modeling and separation.” In Transactions on Machine Learning Research (TMLR), 2023.

# Problem setting

## Separating multiple sources in sequential data

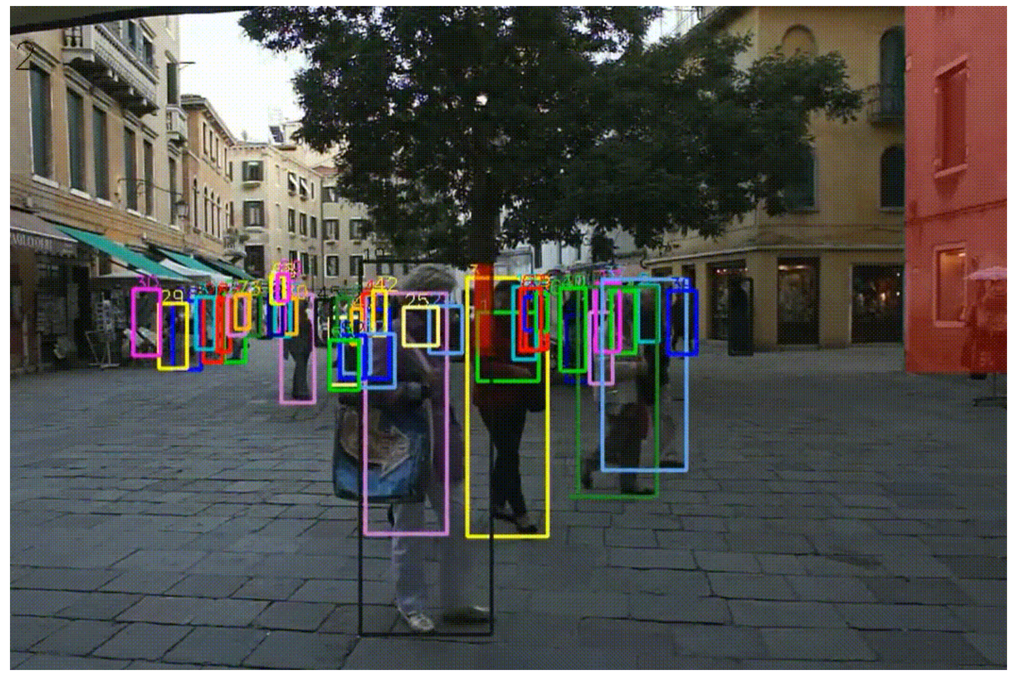



$\mathbf{o}_{1:T,1:K_t}$ $\qquad$ $\mathbf{s}_{1:T,1:N}$

Estimate

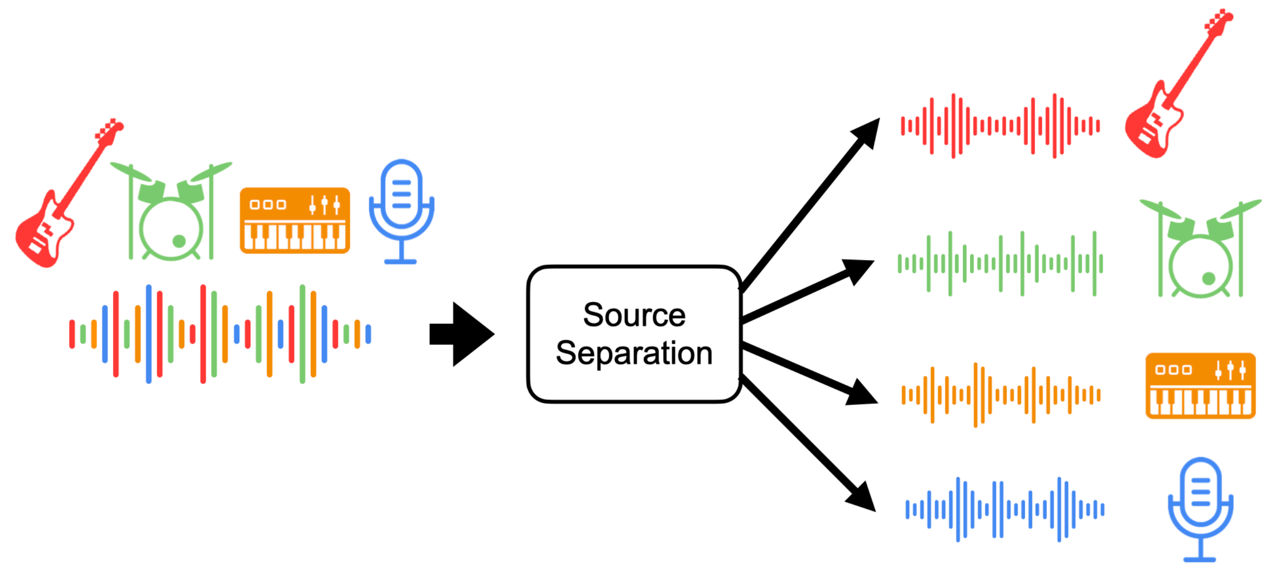
$$P(\mathbf{s}_{1:T,1:N} \mid \mathbf{o}_{1:T,1:K_t})$$

## Application scenarios



Multi-Object Tracking (MOT)

Given a sequence of video, track the objects of interest and assign a unique ID to each of the object.




Single-Channel Audio Source Separation (SC-ASS)

Given a mixture of audio signals, separate different audio sources.

# Proposed solution

## Leveraging Bayesian inference

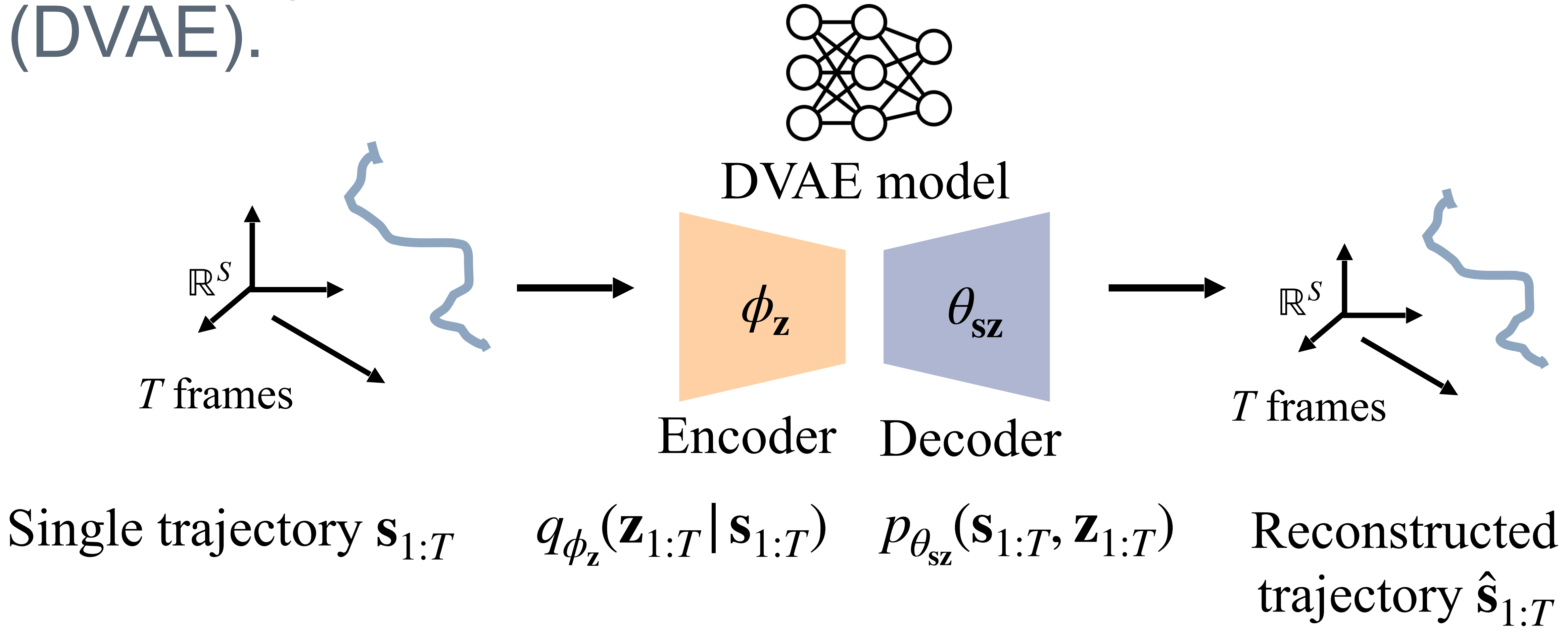
likelihood    prior

$$p_\theta(\mathbf{s} \mid \mathbf{o}) = \frac{p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s}) ds}$$

posterior

marginal likelihood / evidence

- Model $p_\theta(\mathbf{o} \mid \mathbf{s})$ with domain specific knowledge.
- Model $p_\theta(\mathbf{s})$ with a dynamical variational auto-encoder (DVAE).




- Infer $p_\theta(\mathbf{s} \mid \mathbf{o})$ with variational inference methodology.

# Probabilistic model

## Definition of random variables

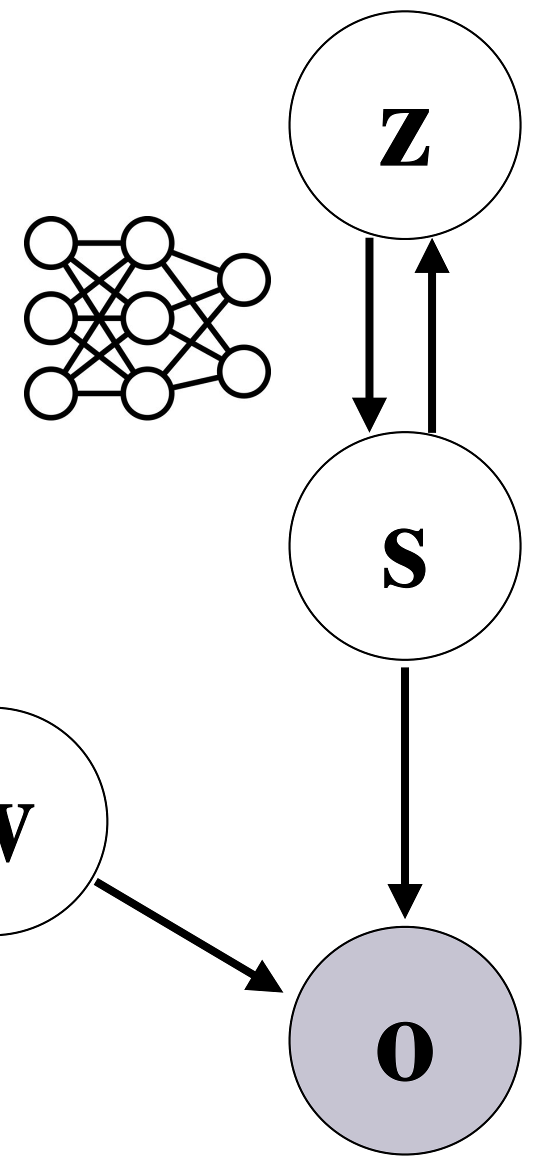
$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: observations.

$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : true source vectors.

$\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times L}$: latent variables of DVAE.

$\mathbf{w} = \{w_{1:T,1:K_t}\} \in \{1, \ldots, N\}^{T \times K_t}$ : discrete assignment variables,

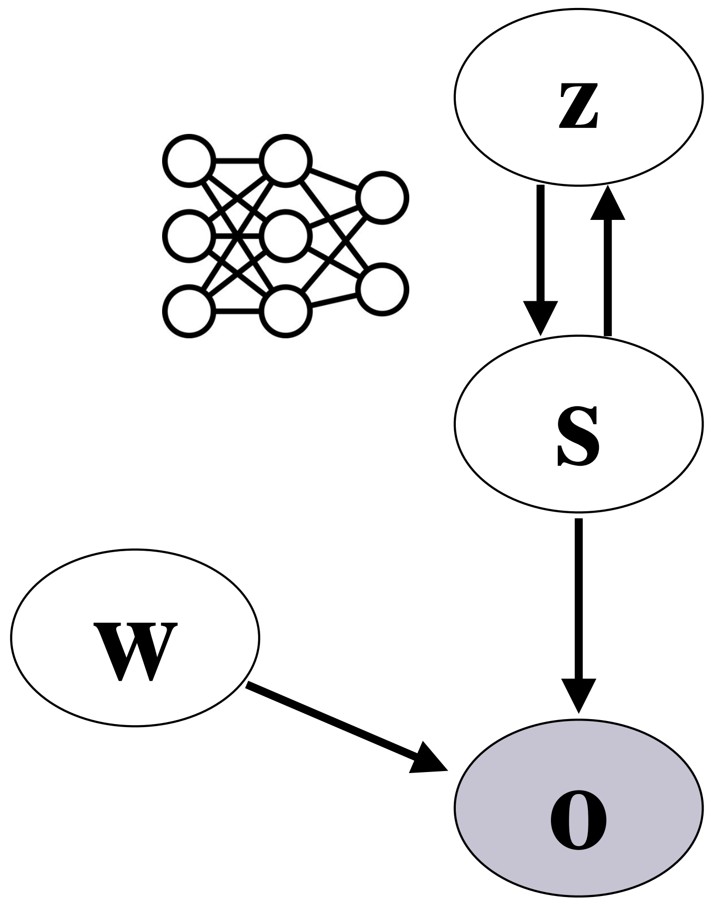
$w_{tk} = n$ indicates the observation $\mathbf{o}_{tk}$ is assigned to source $n$.



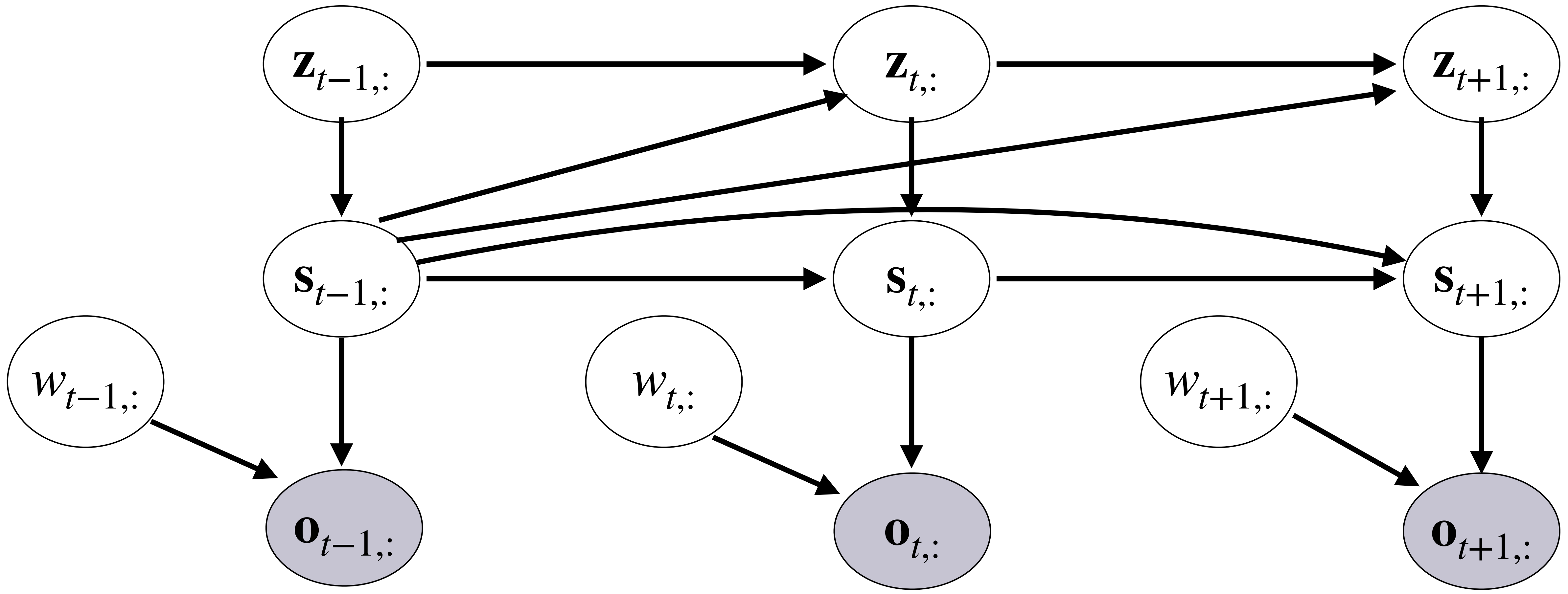

Observed variable: $\mathbf{o}$ Latent variables: $\mathbf{s}$, $\mathbf{z}$, $\mathbf{w}$

Objective: Estimate the posterior distribution $p(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$.

# Probabilistic model




Folded graphical model




Extended graphical model over time frames

**Generative model:** $p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o} \mid \mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_{\mathbf{sz}}}(\mathbf{s}, \mathbf{z})$.

Intractable true posterior distribution $p_\theta(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$.

**Inference model:** factorized approximation $q_\phi(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o}) = q_{\phi_\mathbf{s}}(\mathbf{s} \mid \mathbf{o}) q_{\phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{s}) q_{\phi_\mathbf{w}}(\mathbf{w} \mid \mathbf{o}) \approx p_\theta(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$.

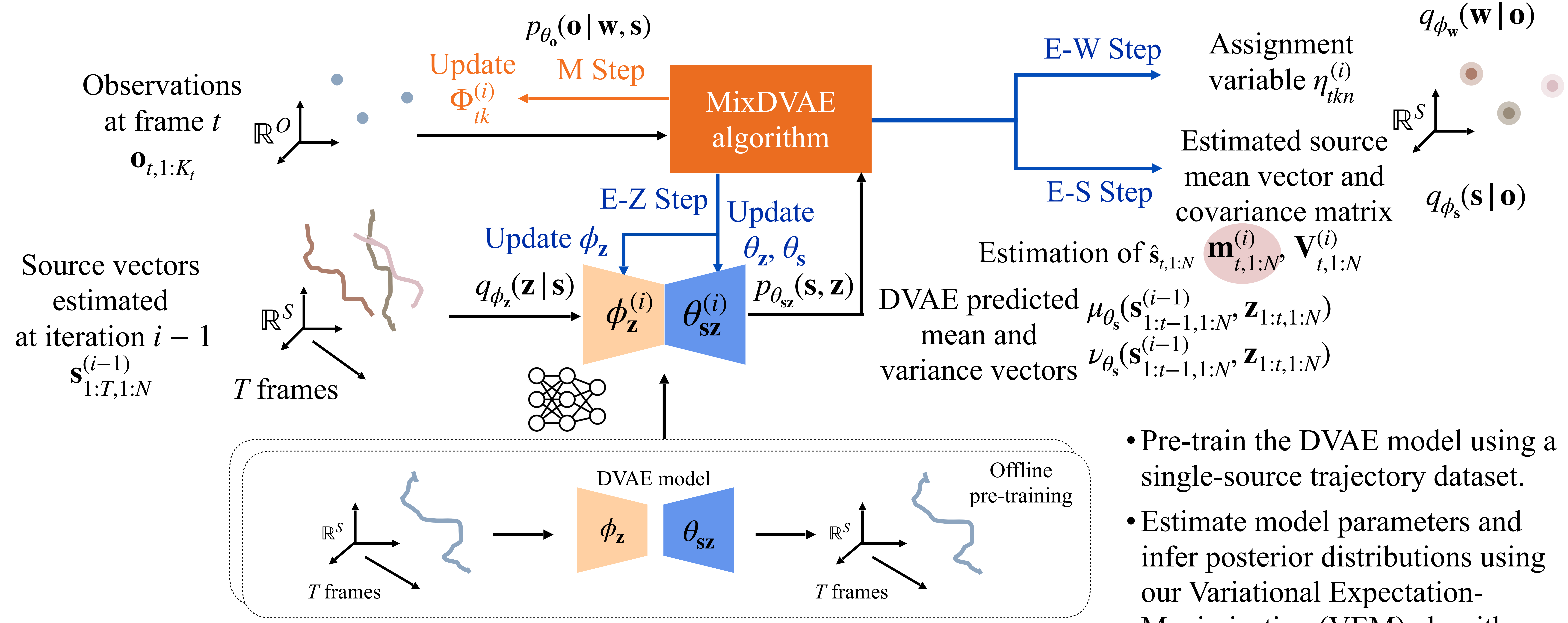
**Optimization:** maximizing the ELBO $\mathscr{L}(\theta, \phi; \mathbf{o}) = \mathbb{E}_{q_\phi(\mathbf{s},\mathbf{z},\mathbf{w}|\mathbf{o})}[\log p_\theta(\mathbf{o}, \mathbf{s}, \mathbf{z}, \mathbf{w}) - \log q_\phi(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})]$.

# MixDVAE algorithm

## Two-step learning framework

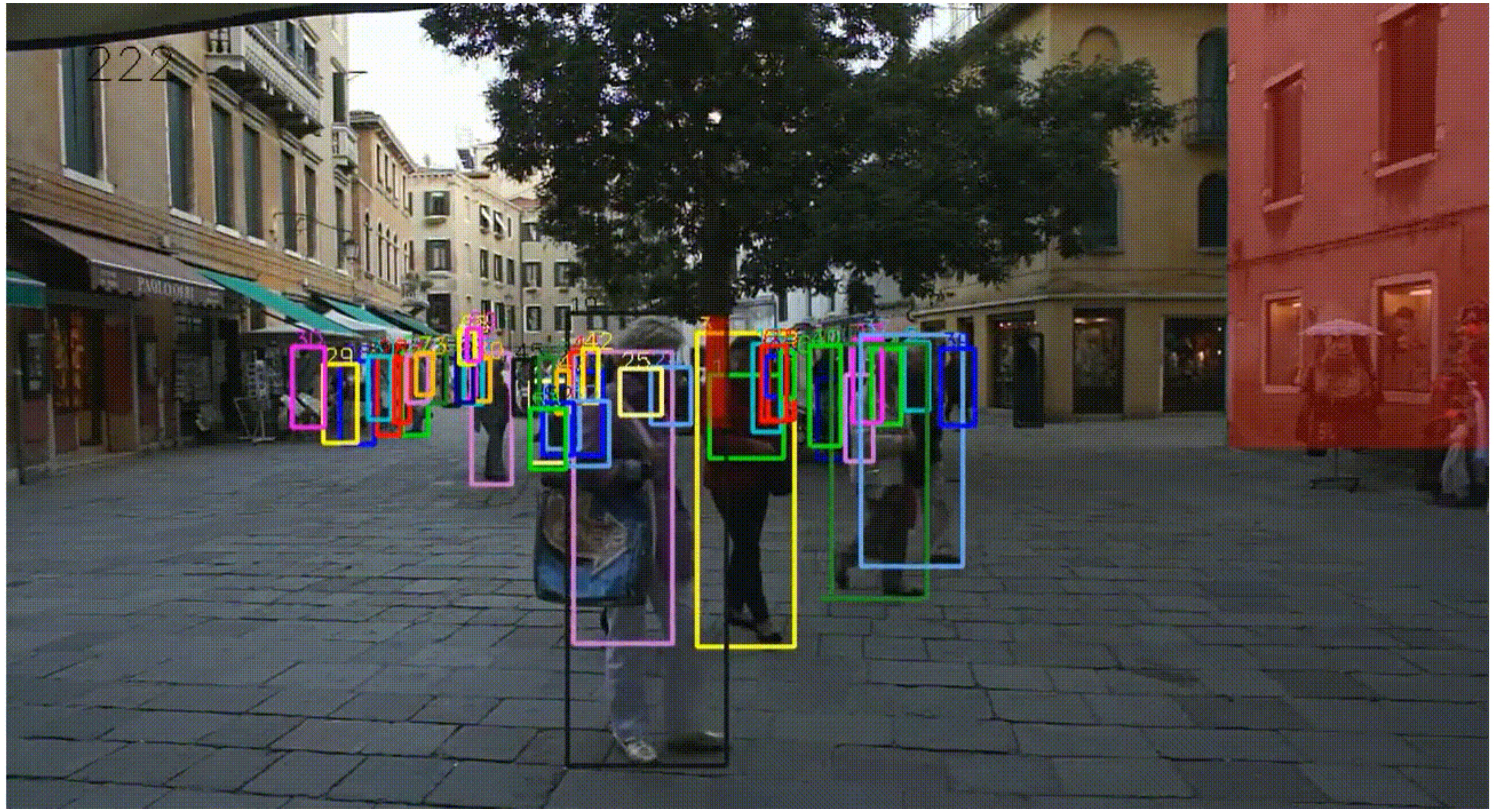
- Pre-train the DVAE model using a single-source trajectory dataset.
- Estimate model parameters and infer posterior distributions using our Variational Expectation-Maximization (VEM) algorithm.

# MixDVAE algorithm




- Pre-train the DVAE model using a single-source trajectory dataset.
- Estimate model parameters and infer posterior distributions using our Variational Expectation-Maximization (VEM) algorithm.

# Applications to MOT

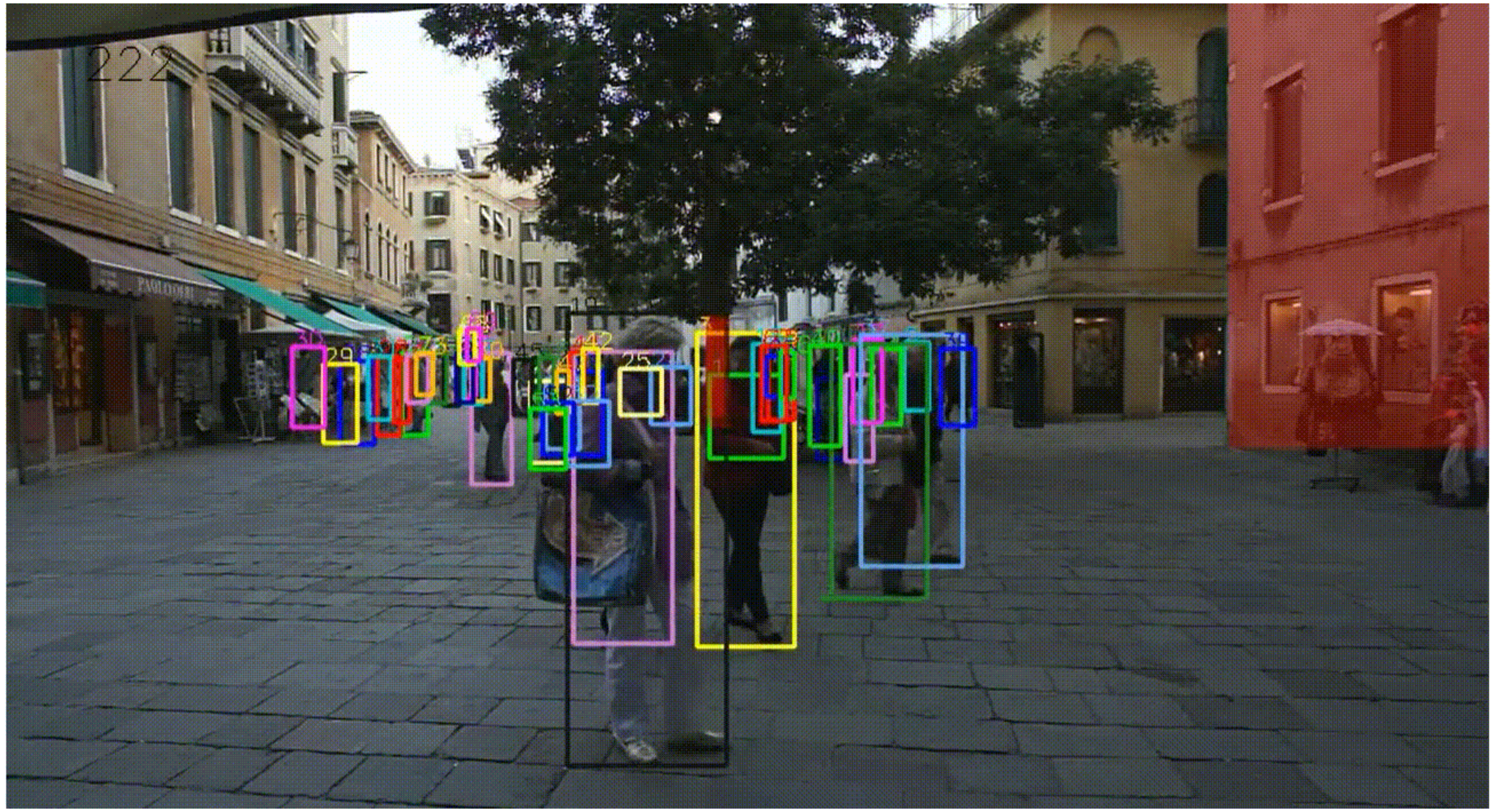


## 4 main sub-tasks in MOT[28,29,30]

- Extracting source observations (detections) at each time frame.
- Modeling the dynamics of the sources.
- Associating observations to sources consistently over time.
- Accounting for birth and death process of source trajectories.

[28] Ba-Ngu Vo, et al. Multitarget Tracking. *Wiley Encyclopedia of Electrical and Electronics Engineering*. 2015.
[29] Wenhan Luo, et al. Multiple object tracking: A literature review. *Artif. Intell.* 2021.
[30] Gioele Ciaparrone, et al. Deep learning in video multi-object tracking: A survey. *Neural Comp.* 2020.

# Applications to MOT




## 4 main sub-tasks in MOT[28,29,30]

- Extracting source observations (detections) at each time frame.
- Modeling the dynamics of the sources.
- Associating observations to sources consistently over time.
- Accounting for birth and death process of source trajectories.

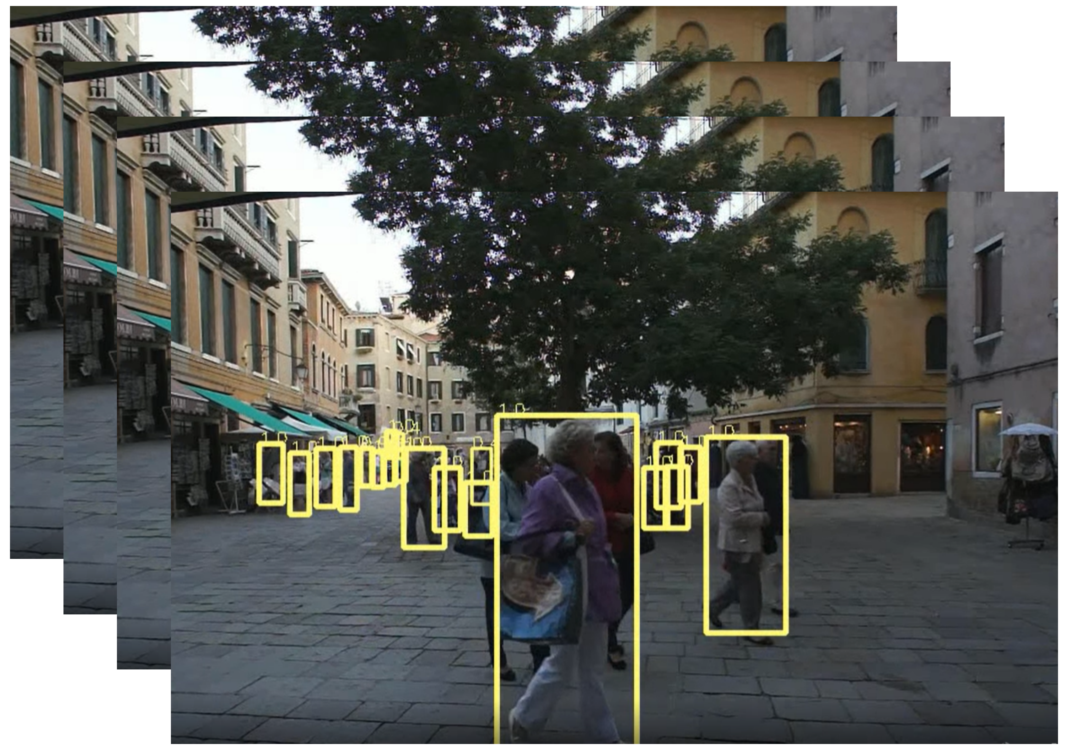
➡ Tracking-by-detection, known number of sources

[28] Ba-Ngu Vo, et al. Multitarget Tracking. *Wiley Encyclopedia of Electrical and Electronics Engineering*. 2015.
[29] Wenhan Luo, et al. Multiple object tracking: A literature review. *Artif. Intell.* 2021.
[30] Gioele Ciaparrone, et al. Deep learning in video multi-object tracking: A survey. *Neural Comp.* 2020.

# Probabilistic model of MOT

## Definition of random variables

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: coordinates of detection bounding boxes.



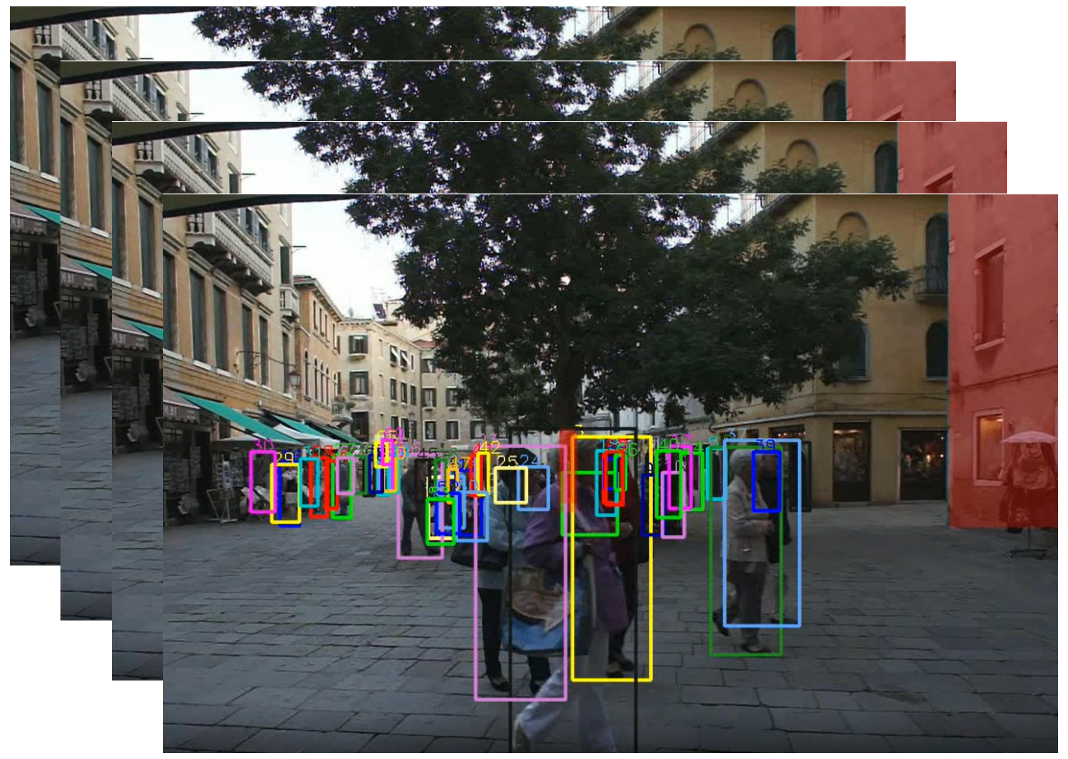
$\mathbf{o}$

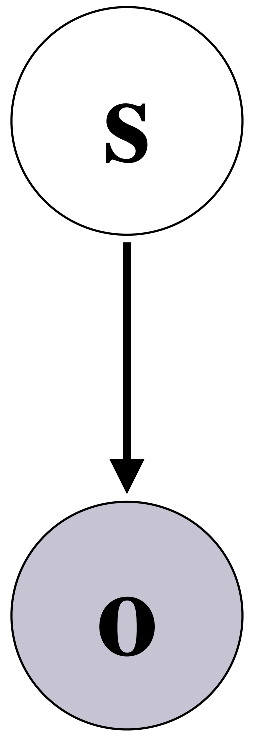
# Probabilistic model of MOT

## Definition of random variables

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: coordinates of detection bounding boxes.

$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : true coordinates of sources.





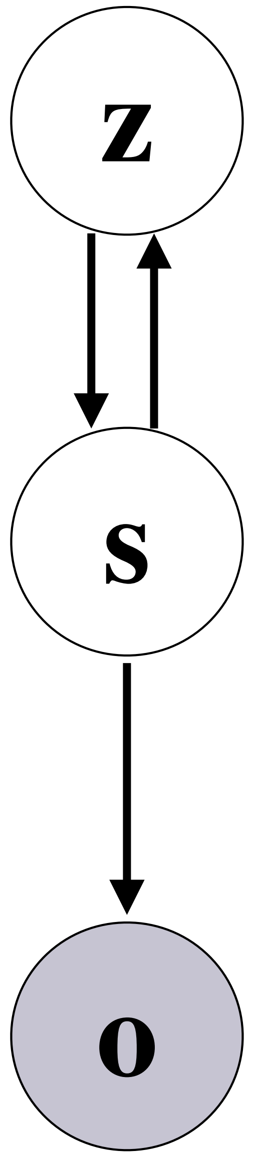
# Probabilistic model of MOT

## Definition of random variables

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T\times K_t \times O}$: coordinates of detection bounding boxes.

$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T\times N \times S}$ : true coordinates of sources.

$\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T\times N \times L}$: latent variables of DVAE.



# Probabilistic model of MOT

## Definition of random variables

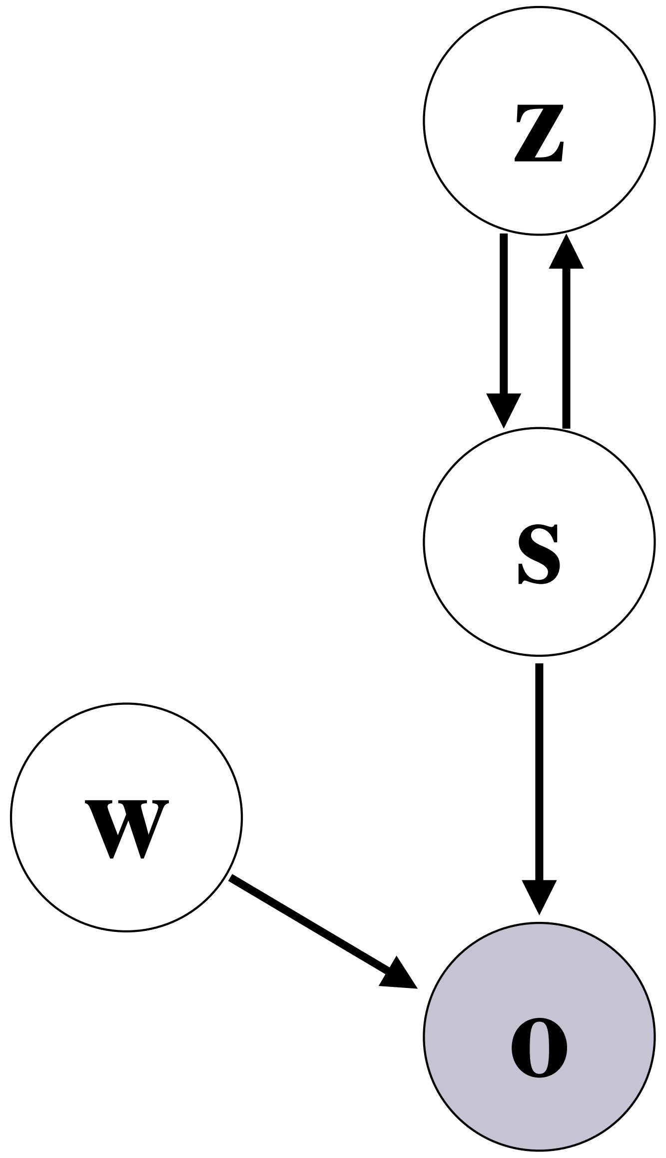
$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: coordinates of detection bounding boxes.

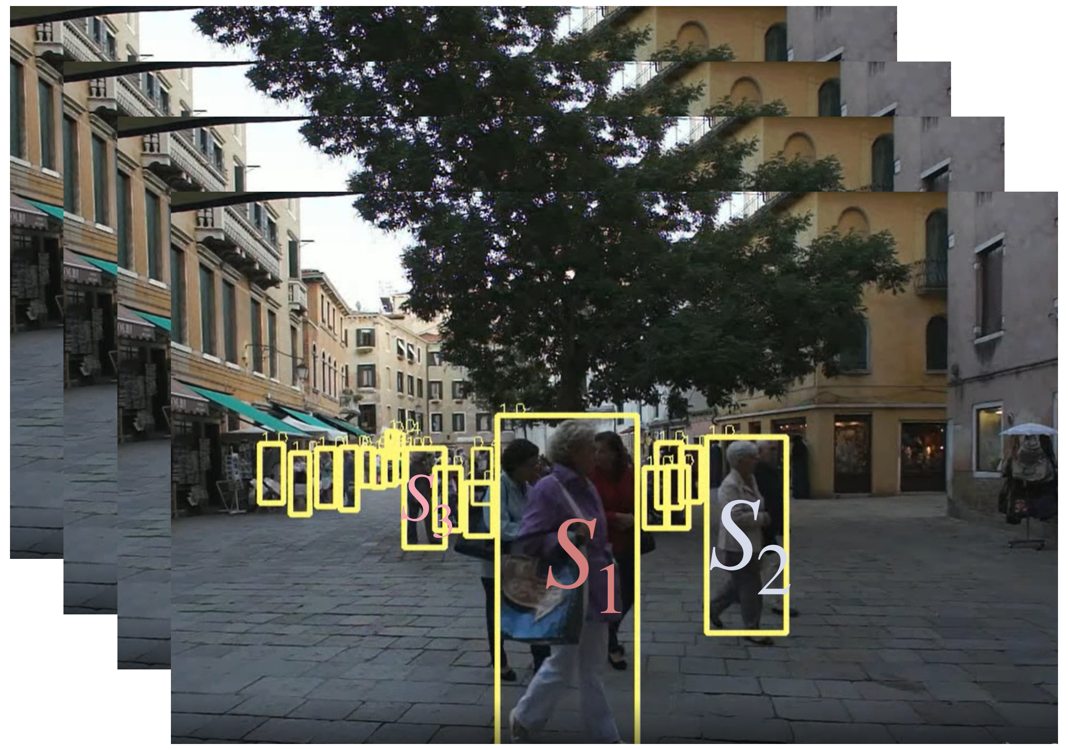
$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : true coordinates of sources.

$\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times L}$: latent variables of DVAE.

$\mathbf{w} = \{w_{1:T,1:K_t}\} \in \{1,\ldots,N\}^{T \times K_t}$ : discrete assignment variables,

$w_{tk} = n$ indicates the detection $\mathbf{o}_{tk}$ is assigned to source $n$.



# Experimental settings

## Datasets

### DVAE pre-training

A synthetic single-source motion trajectories dataset

### Unsupervised MOT Evaluation

MOT17-3T dataset created from the MOT17[31] training set:

- Subsequences of length $T$ ($T = 60, 120, 300$ frames are tested)
- 3 tracking sources per test data sample

## Baselines

ArTIST[32] (LSTM-based supervised method), VKF[33] (linear filtering method), Deep AR (LSTM-based filtering method)

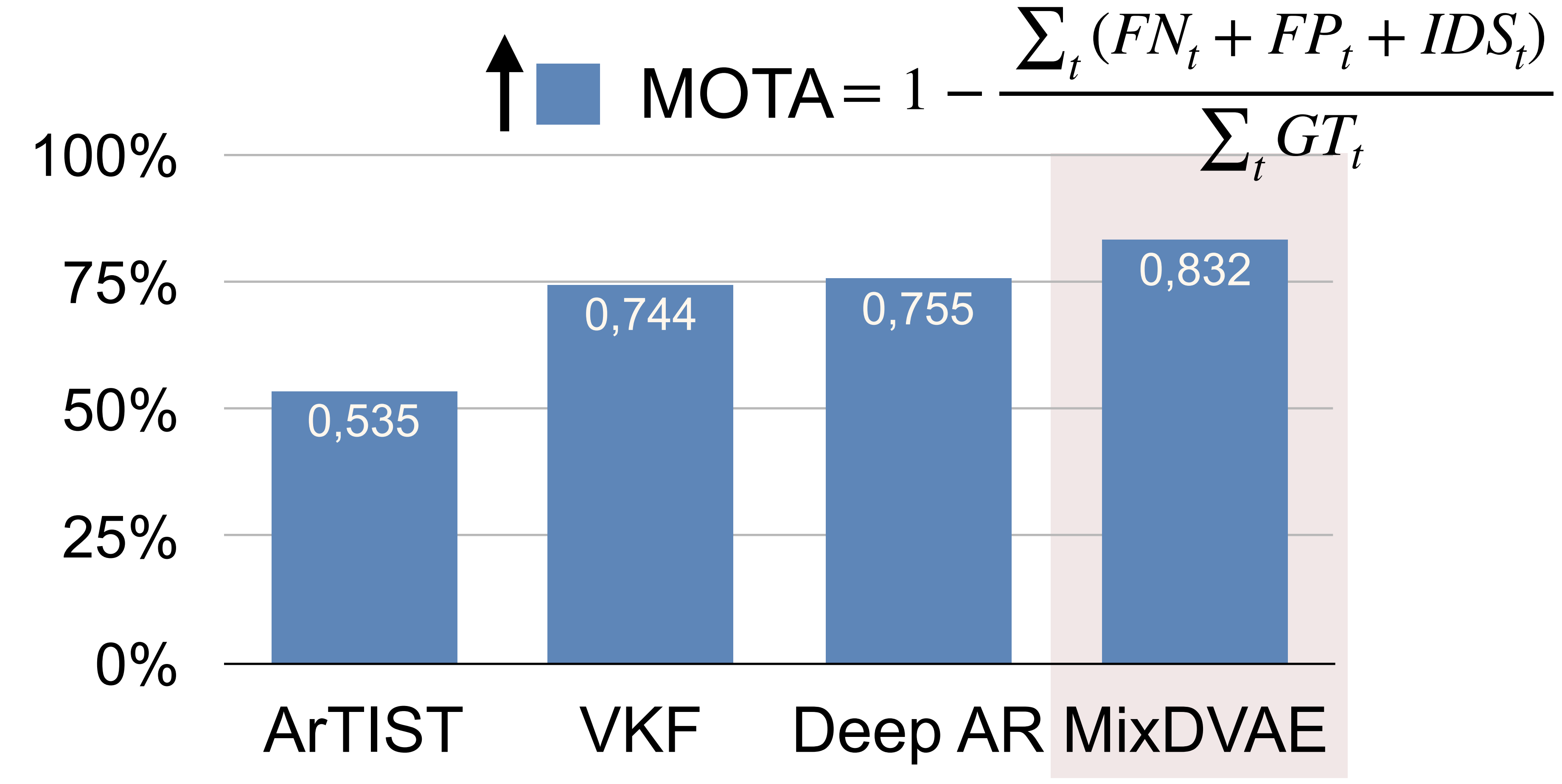
## Evaluation metrics[34,35]

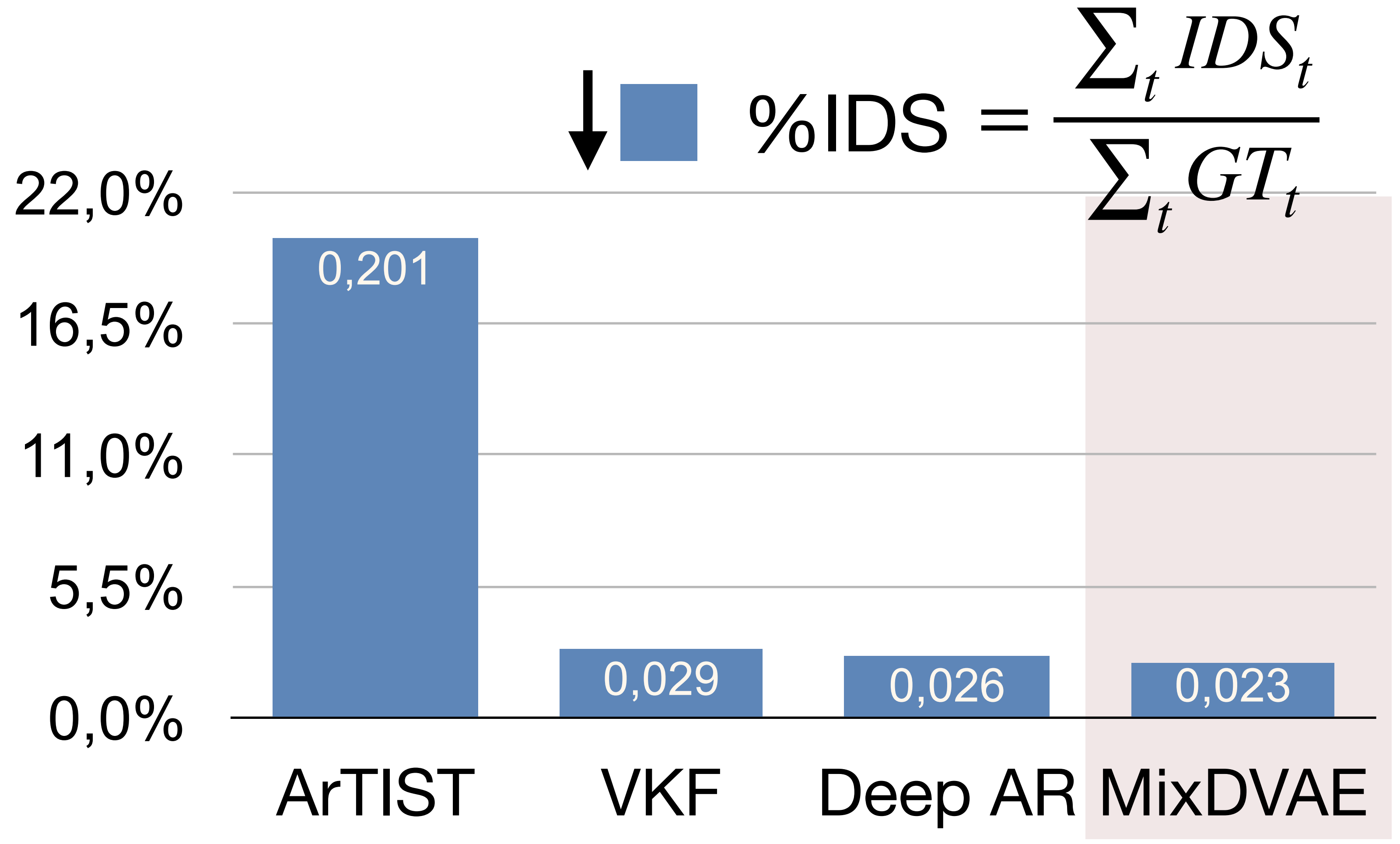
Multi-object tracking accuracy (MOTA), number of identity switches (IDS), false positives (FP), false negatives (FN)

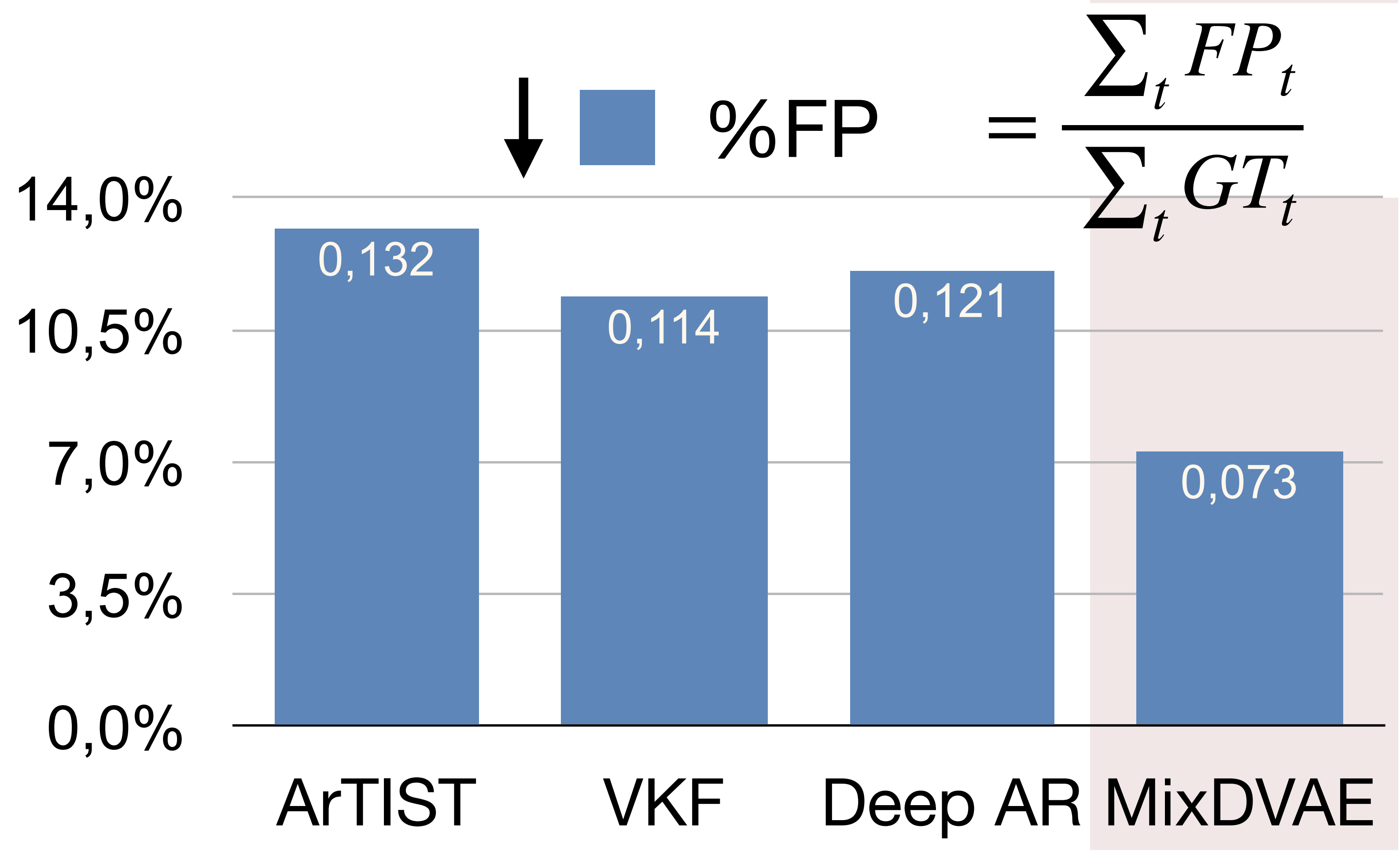
[31] Patrick Dendorfer, et al. MOTChallenge: A benchmark for single-camera multiple target tracking. *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. 2021.
[32] Fatemeh Saleh, et al. Probabilistic tracklet scoring and inpainting for multiple object tracking. *Proc. IEEE Int. Conf. Computer Vision Pattern Recogn. (CVPR)*. 2021.
[33] Yutong Ban, et al. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021.
[34] Keni Bernardin, et al. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.* 2008
[35] Ergys Ristani, et al. Performance measures and a data set for multi-target, multi-camera tracking. *Proc. Europ. Conf. Computer Vision (ECCV)*. 2016.

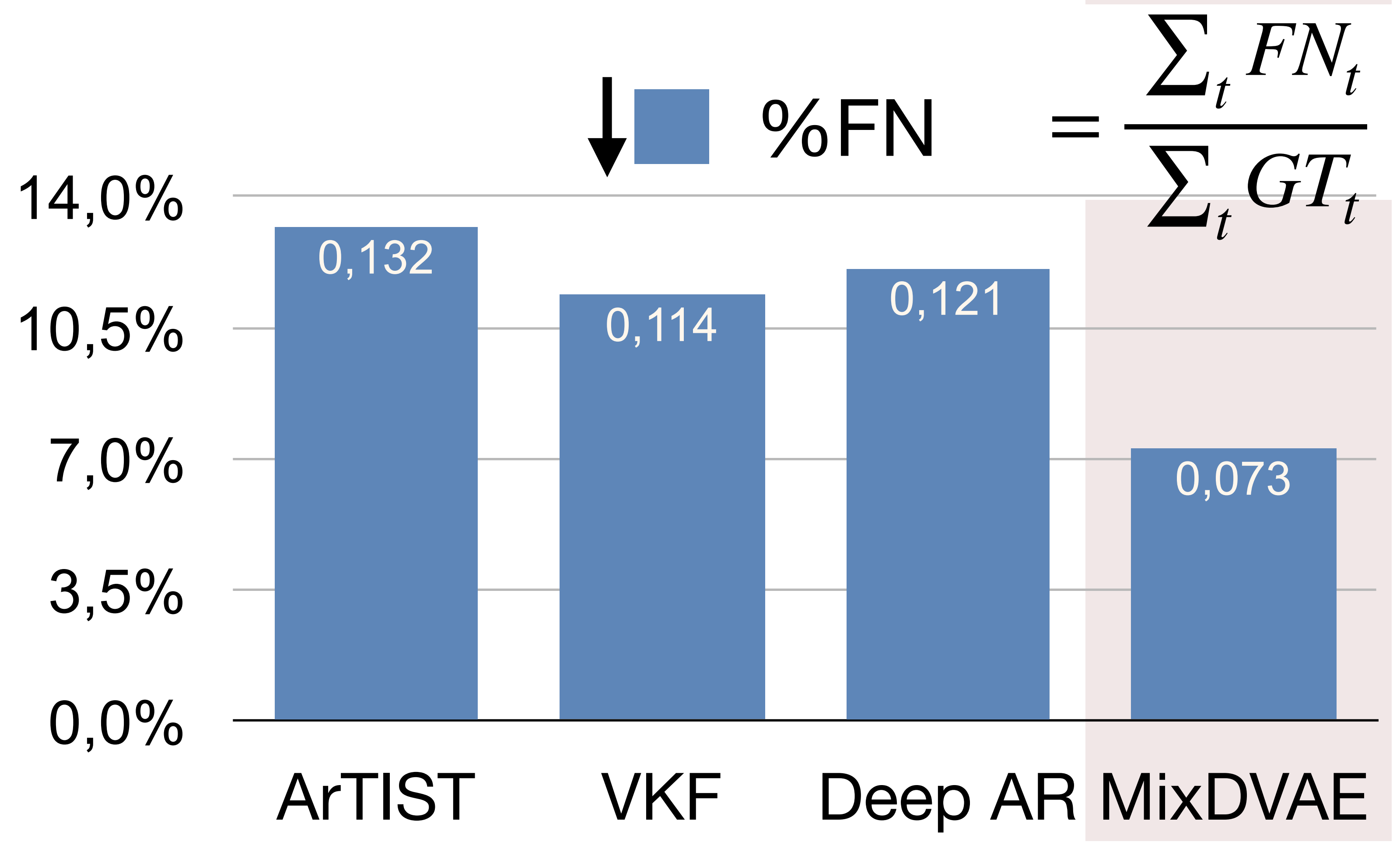
# Quantitative analysis

**Evaluation on long sequences ($T = 300$).**



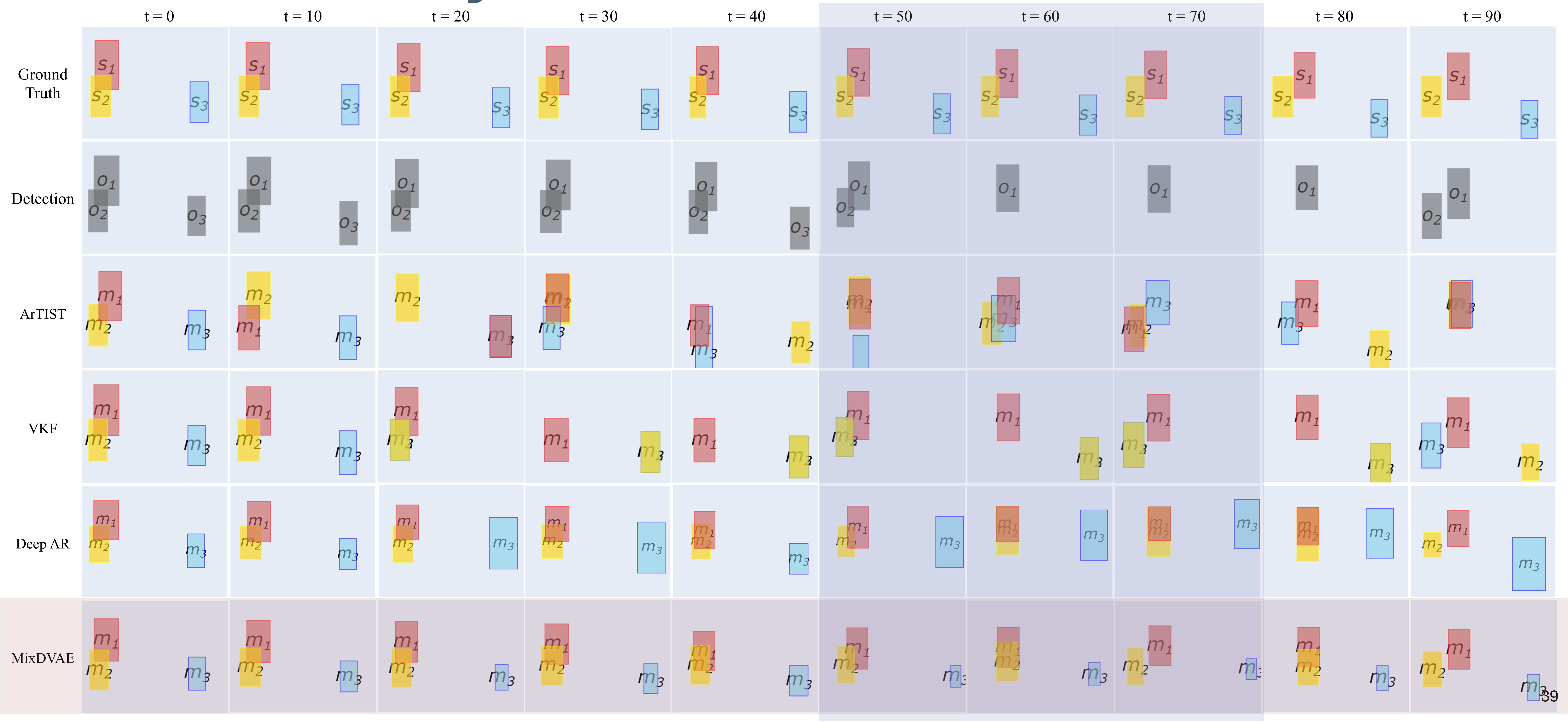



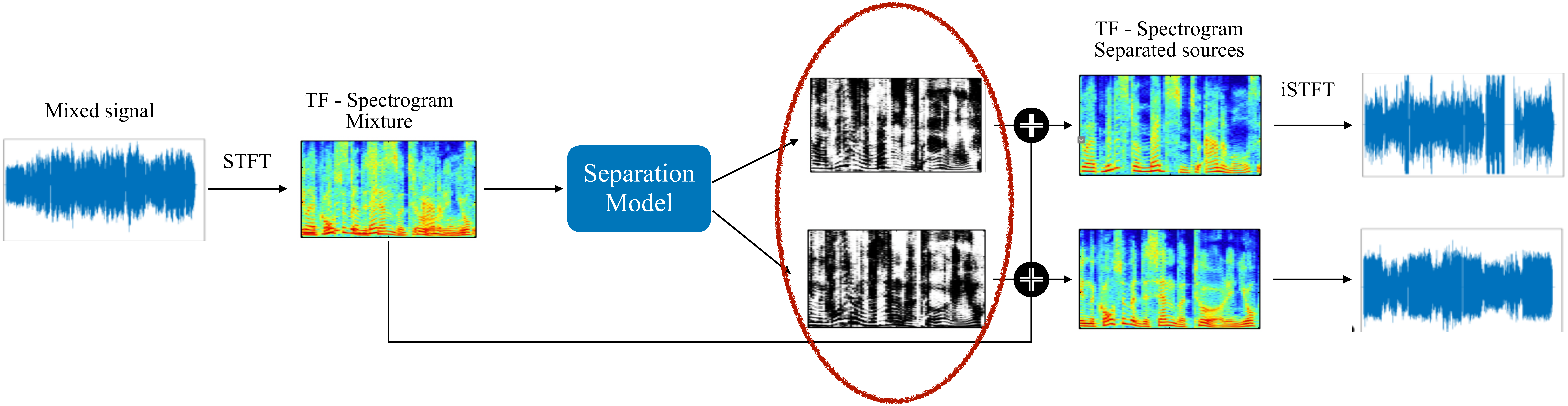
# Qualitative analysis

Robust tracking with frequent occlusions.



# Applications to SC-ASS

## Mask-based method[36,37,38,39]



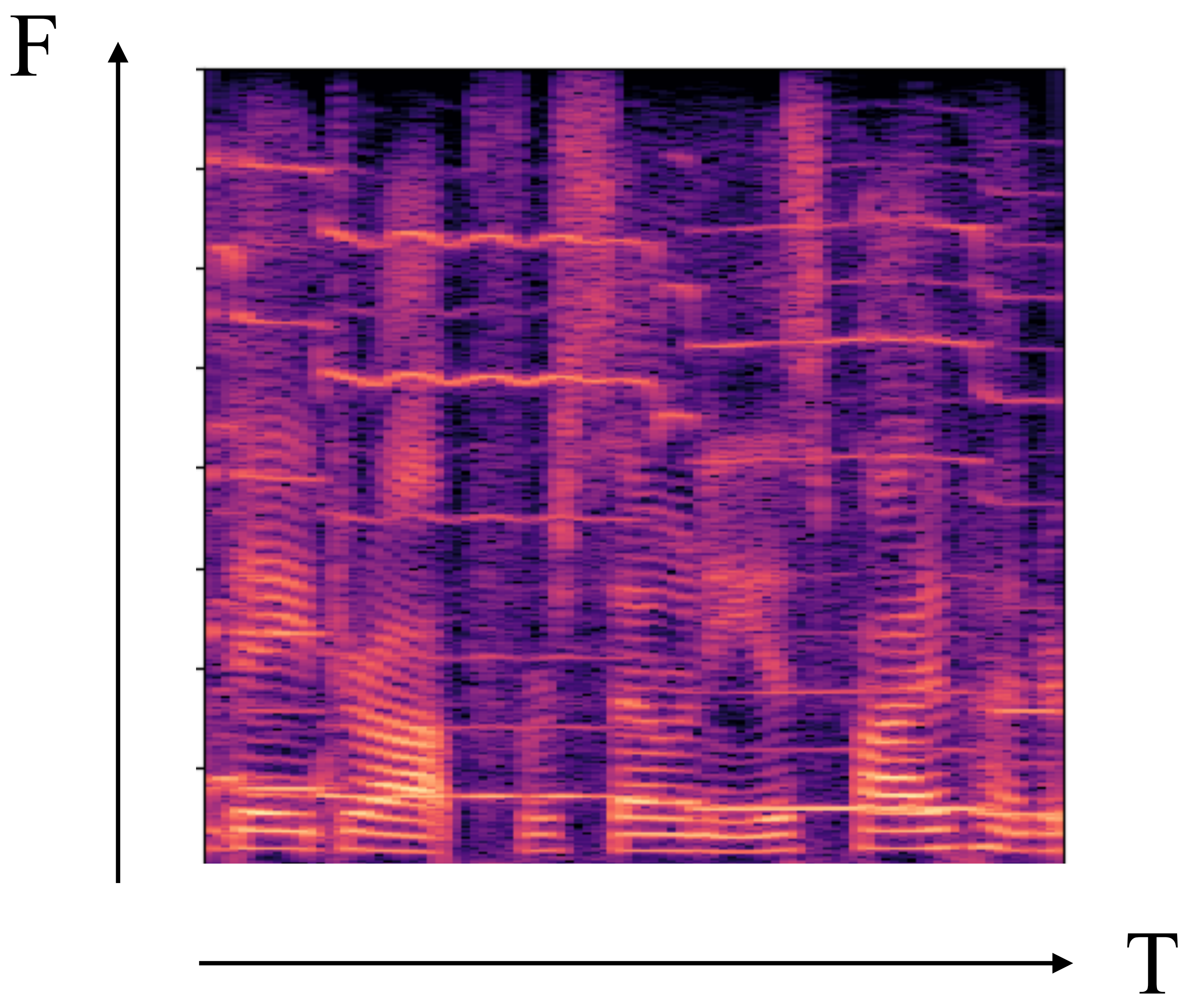

Key question: how to obtain the masks?

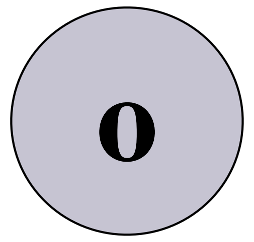
[36] Emmanuel Vincent, et al. Audio Source Separation and Speech Enhancement. 2018.
[37] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via timefrequency masking. *IEEE Trans. Signal Process.* 2004.
[38] Dorothea Kolossa, et al. Nonlinear postprocessing for blind speech separation. *Independent Component Analysis and Blind Signal Separation.* 2004.
[39] DeLiang Wang and Guy J. Brown. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. 2006.

# Probabilistic model of SC-ASS

## Definition of random variables

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: STFT spectrogram of the observed mixture signal.



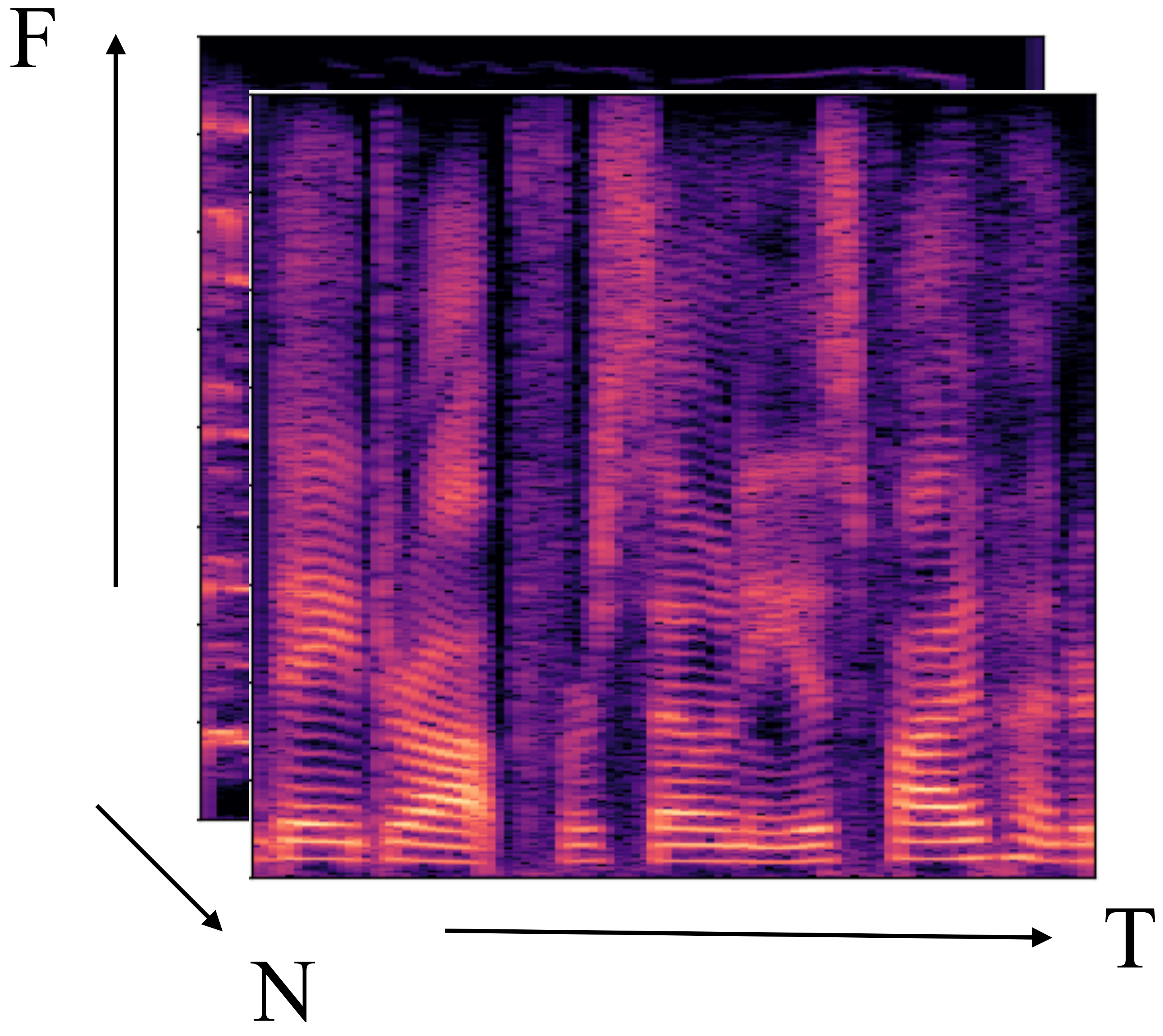



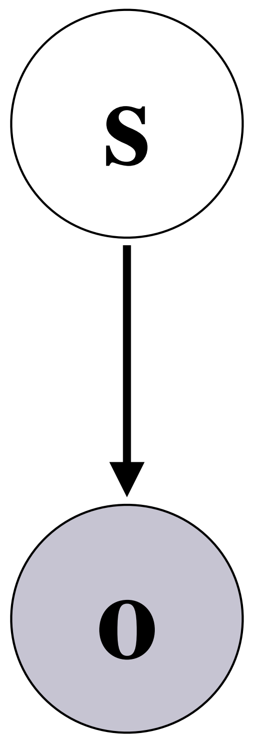
# Probabilistic model of SC-ASS

## Definition of random variables

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: STFT spectrogram of the observed mixture signal.

$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : STFT spectrograms of $N$ sources.

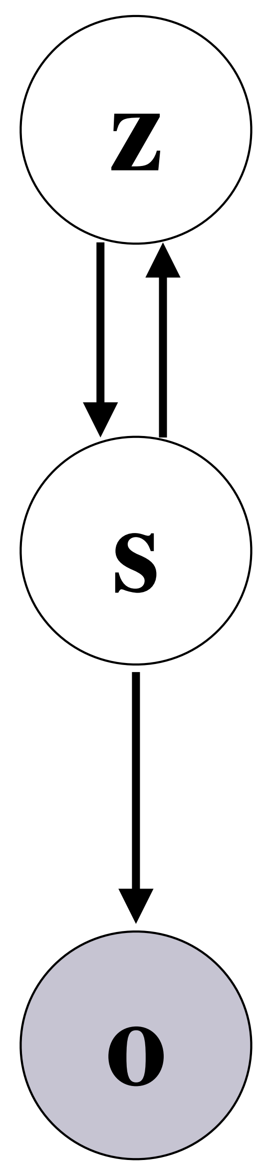


# Probabilistic model of SC-ASS

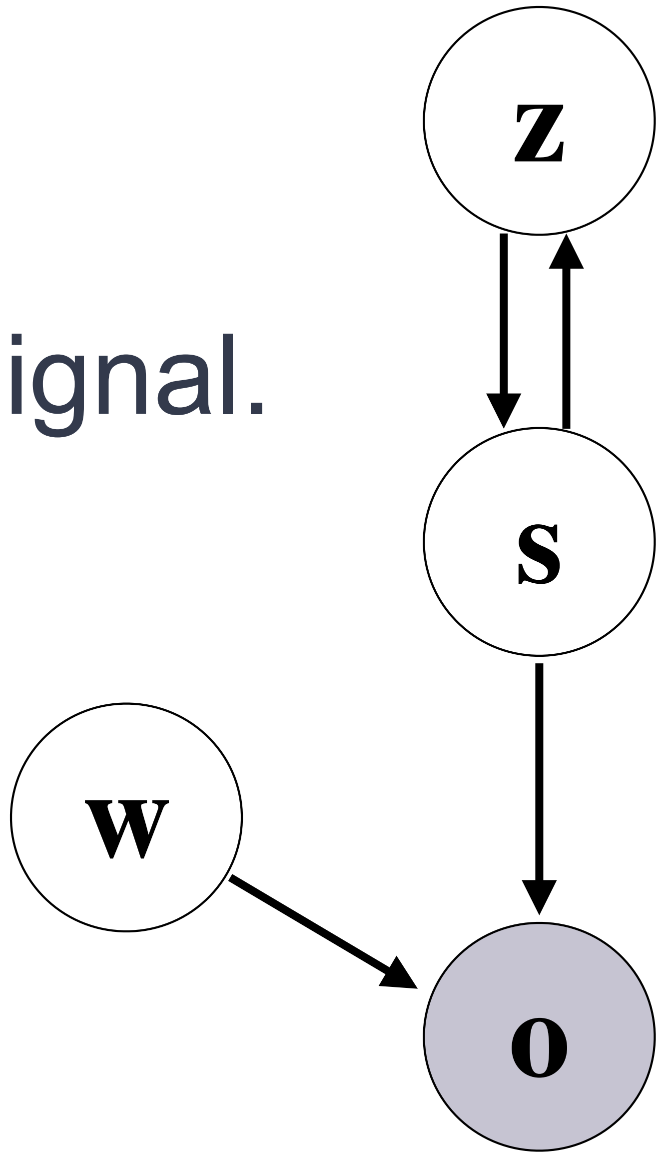
## Definition of random variables

- $\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: STFT spectrogram of the observed mixture signal.
- $\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : STFT spectrograms of $N$ sources.
- $\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times L}$: latent variables of DVAE.



# Probabilistic model of SC-ASS

## Definition of random variables



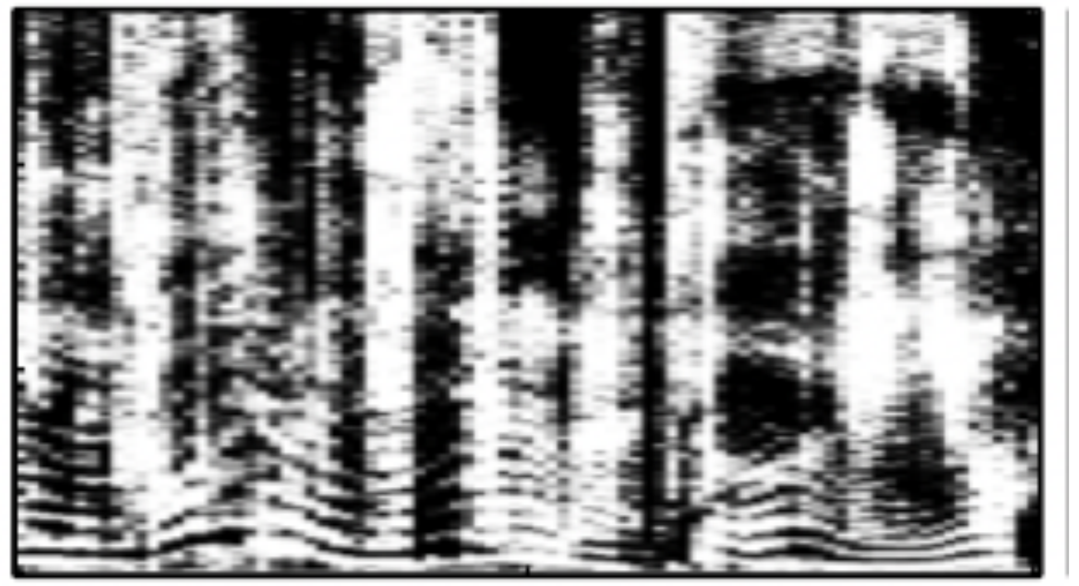

$\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times O}$: STFT spectrogram of the observed mixture signal.

$\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times S}$ : STFT spectrograms of $N$ sources.

$\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times L}$: latent variables of DVAE.

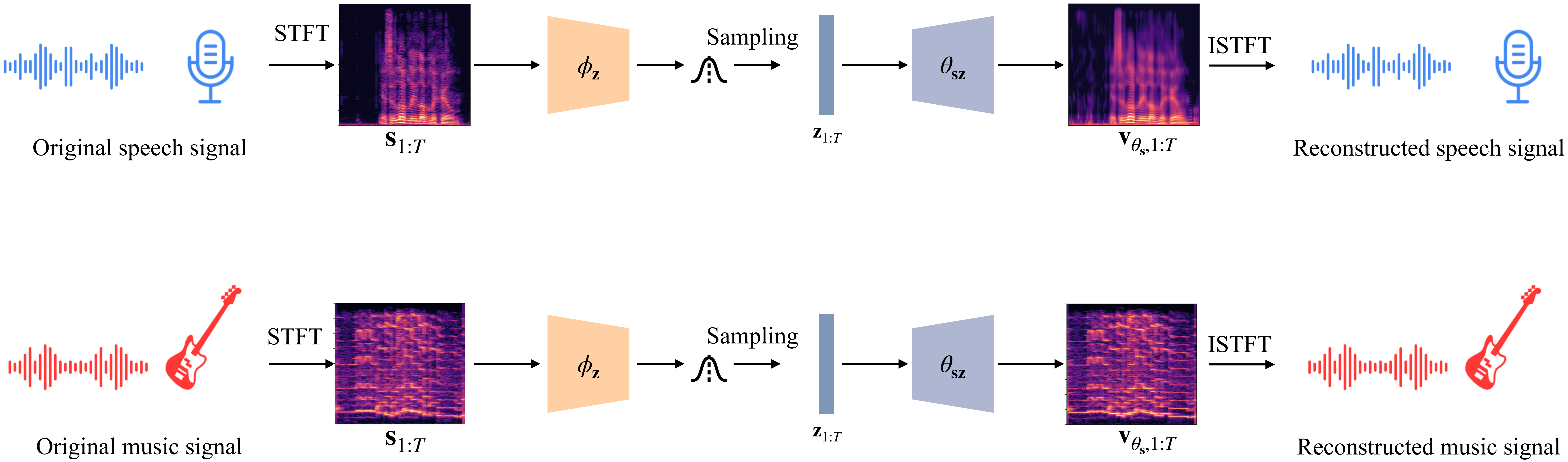
$\mathbf{w} = \{w_{1:T,1:K_t}\} \in \{1,\ldots,N\}^{T \times K_t}$ : discrete assignment variables,

$w_{tk} = n$ indicates the mixture signal at TF bin [t, f] $o_{t,f}$ is assigned to source $n$.



# Applications to SC-ASS

## Pre-train a DVAE model on each single audio source signal



# Experimental settings

## Datasets

### DVAE pre-training

- Wall Street Journal (WSJ0) dataset[40]
- Chinese Bamboo Flute (CBF) dataset[41]

### Weakly-supervised SC-ASS Evaluation

Mixture signals created from the WSJ0 and CBF test sets with different speech-to-music ratios and three different sequence lengths ($T = 50, 100, 300$)

## Baselines

VKF (linear filtering method), Deep AR (LSTM-based filtering method), MixIT[42] (DL-based unsupervised method), Vanilla NMF[43,44], temporal NMF[45] (statistical method)

## Evaluation metrics

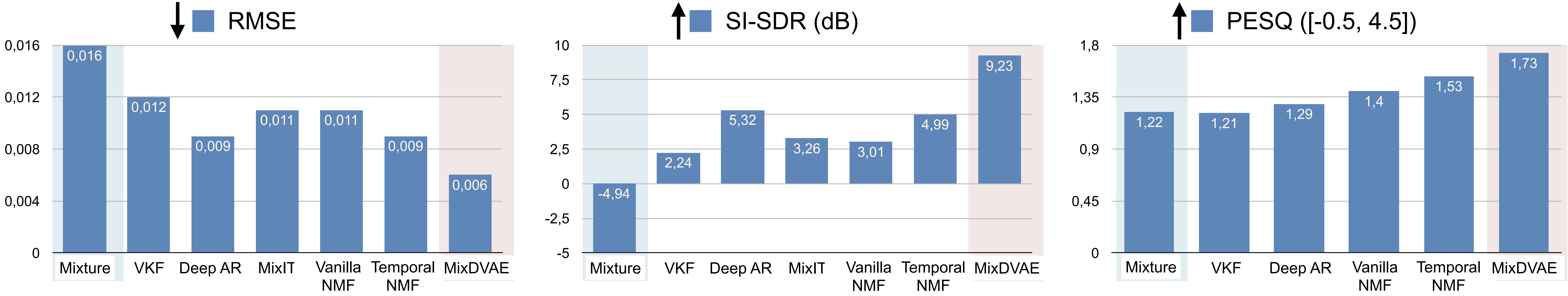
Root mean squared error (RMSE), scale-invariant signal-to-distortion ratio (SI-SDR)[46] (in dB), perceptual evaluation of speech quality (PESQ)[47] (in $[-0.5, 4.5]$).

[40] John S. Garofolo, et al. CSR-I (WSJ0) Sennheiser LDC93S6B. *Philadelphia: Linguistic Data Consortium*. 1993.
[41] Changhong Wang, et al. Adaptive scattering transforms for playing technique recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 2022.
[42] Scott Wisdom, et al. Unsupervised sound separation using mixture invariant training. *Advances in Neural Inform. Process. Systems (NeurIPS)*. 2020.
[43] Cédric Févotte, et al. Single-Channel Audio Source Separation with NMF: Divergences, Constraints and Algorithms. 2018.
[44] Alexey Ozerov, et al. Coding-Based Informed Source Separation: Nonnegative Tensor Factorization Approach. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 2013.
[45] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio, Speech, Lang. Process*. 2007.
[46] Jonathan Le Roux, et al. SDR–Half-baked or well done? *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2019.
[47] Antony Rix, et al. Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2001.

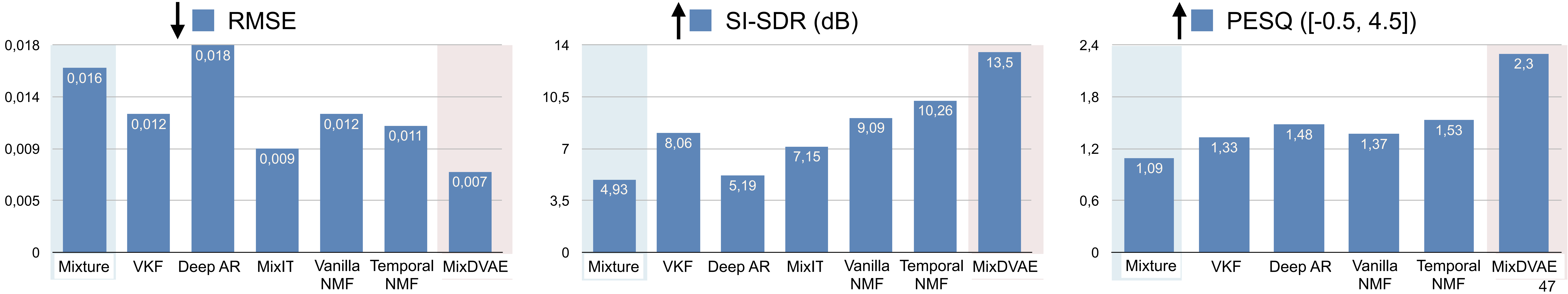
# Quantitative analysis

**Evaluation on short sequences ($T = 50$).**

## Speech




## Chinese bamboo flute

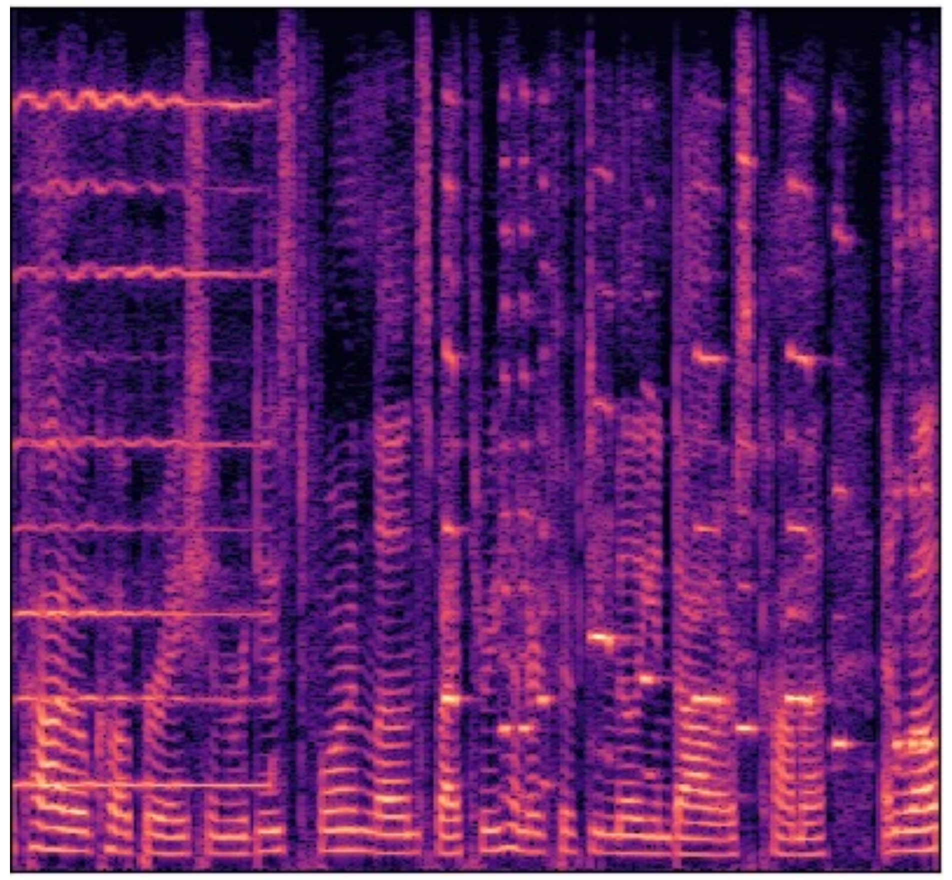


# Qualitative analysis

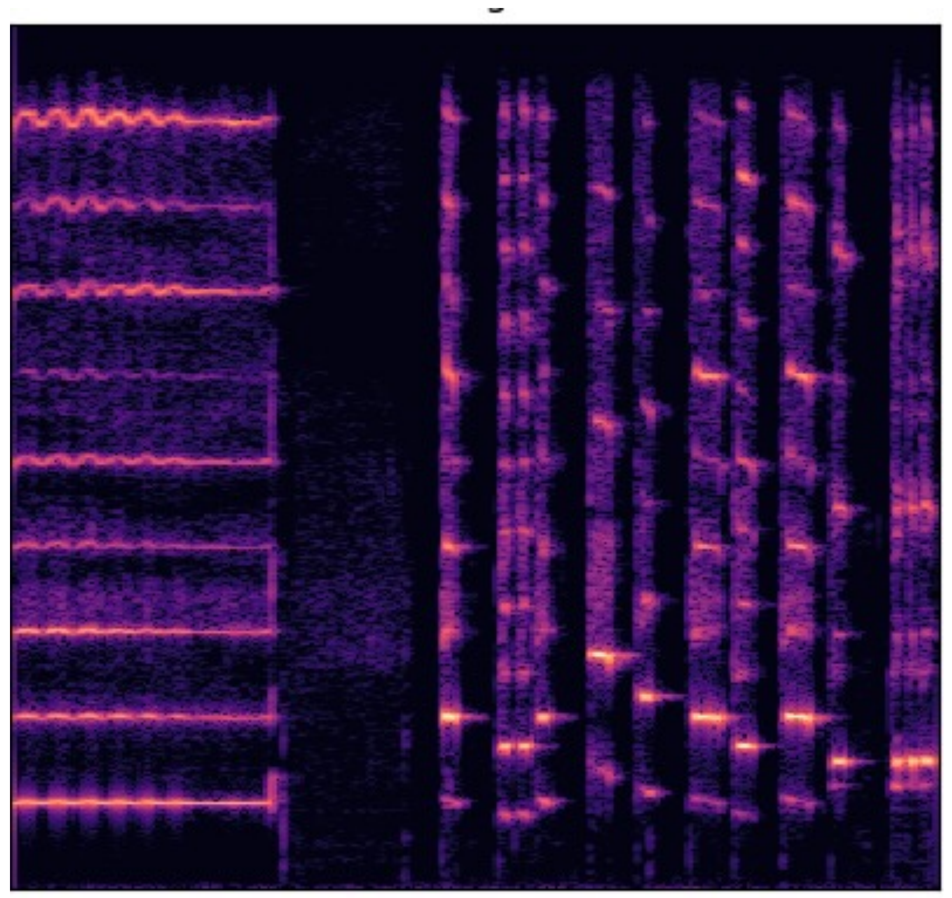
Ground Truth

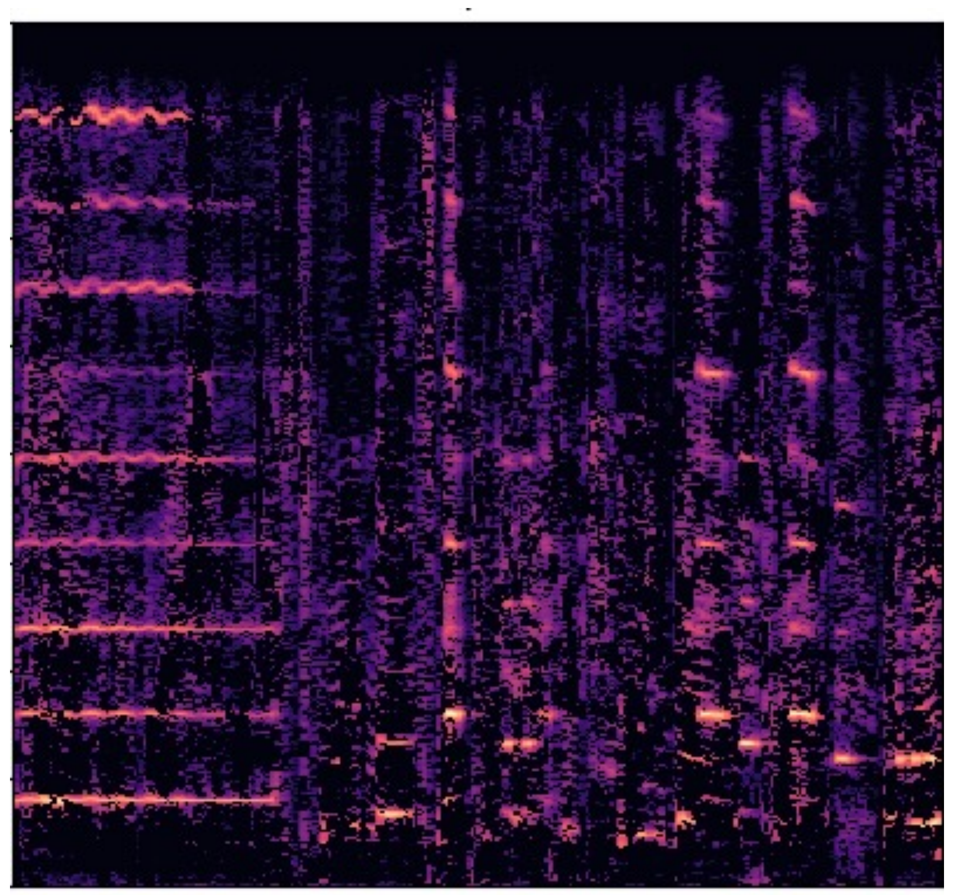
MixDVAE Separation

Mixture

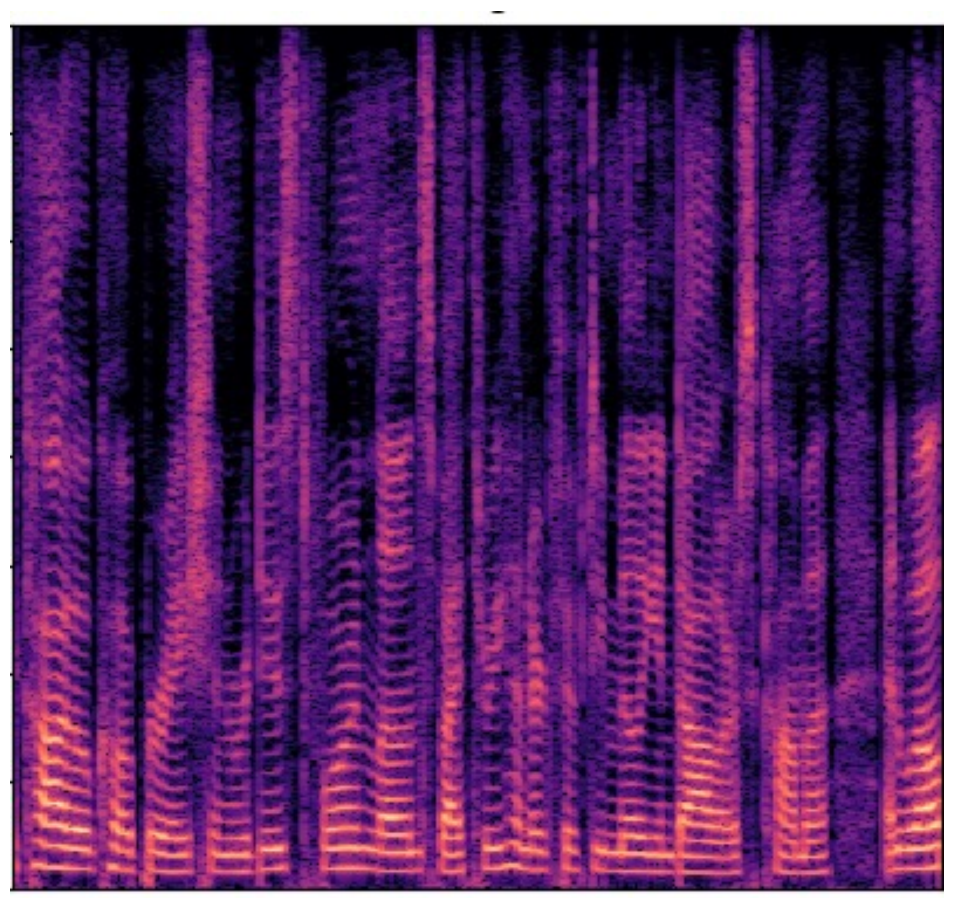


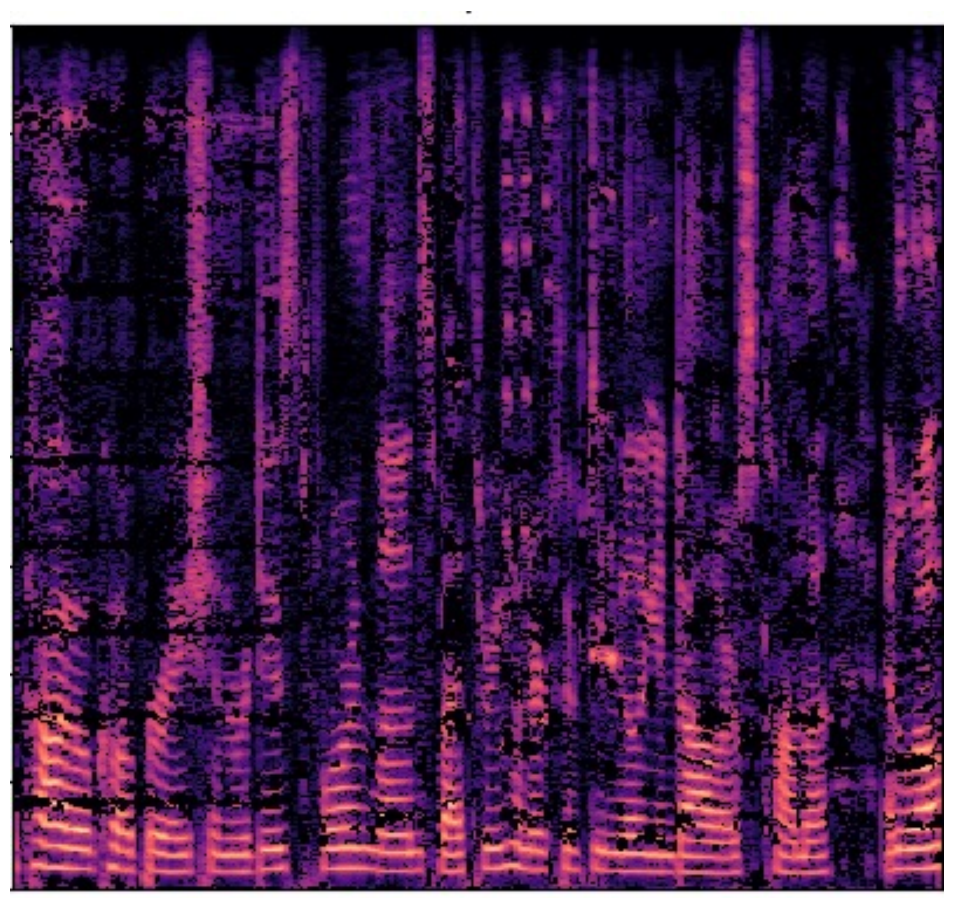
Chinese bamboo flute





Speech





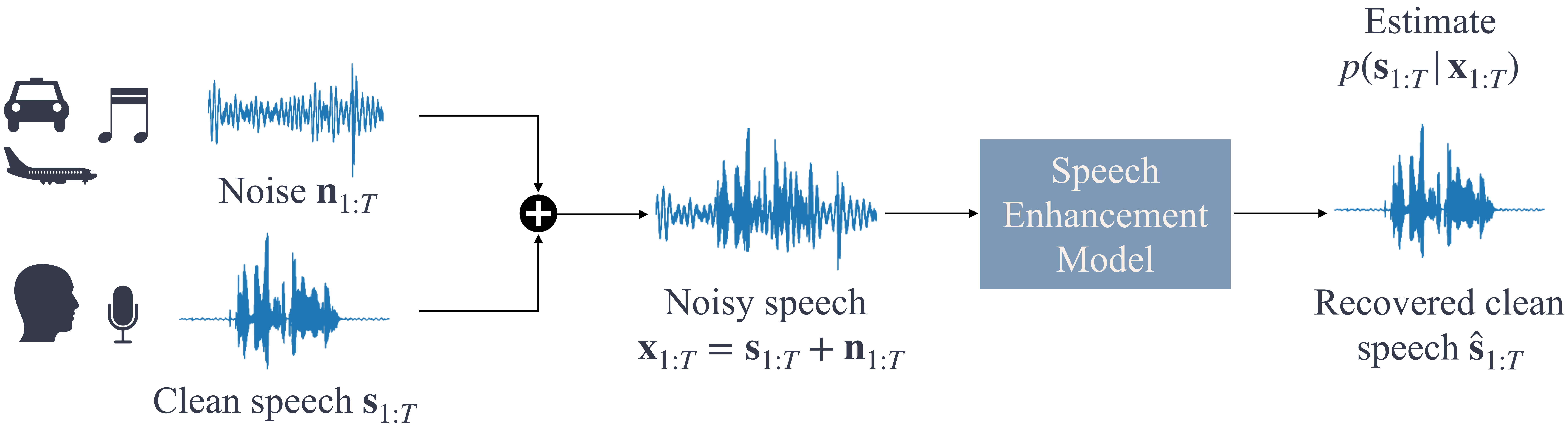
# Part 2

# Unsupervised speech enhancement with deep dynamical probabilistic generative models

Xiaoyu Lin, Simon Leglaive, Laurent Girin, and Xavier Alameda-Pineda. “Unsupervised speech enhancement with deep dynamical generative speech and noise models.” In Proceedings Interspeech Conference, 2023.

# Speech enhancement under additive noise assumption

**Objective:** recover clean speech from noisy speech signals.



# Speech enhancement with Bayesian inference

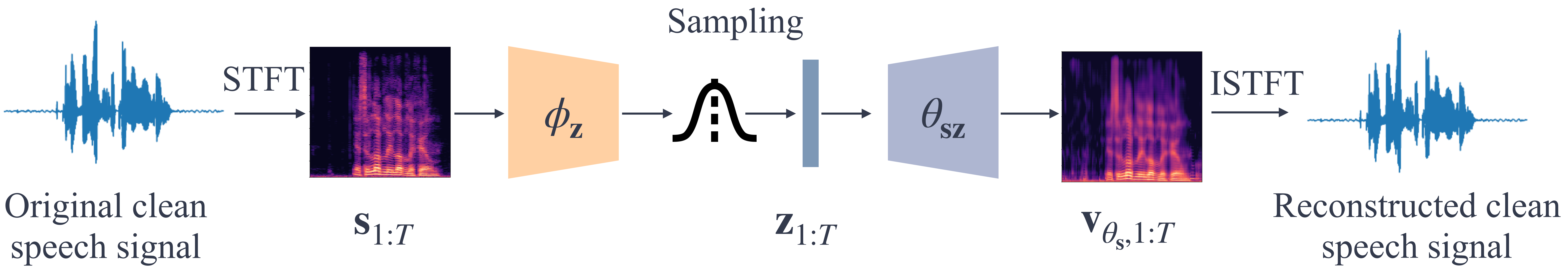
likelihood — prior

$$p_\theta(\mathbf{s}\,|\,\mathbf{x}) = \frac{p_\theta(\mathbf{x}\,|\,\mathbf{s})p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{x}\,|\,\mathbf{s})p_\theta(\mathbf{s})ds}$$

posterior

marginal likelihood / evidence

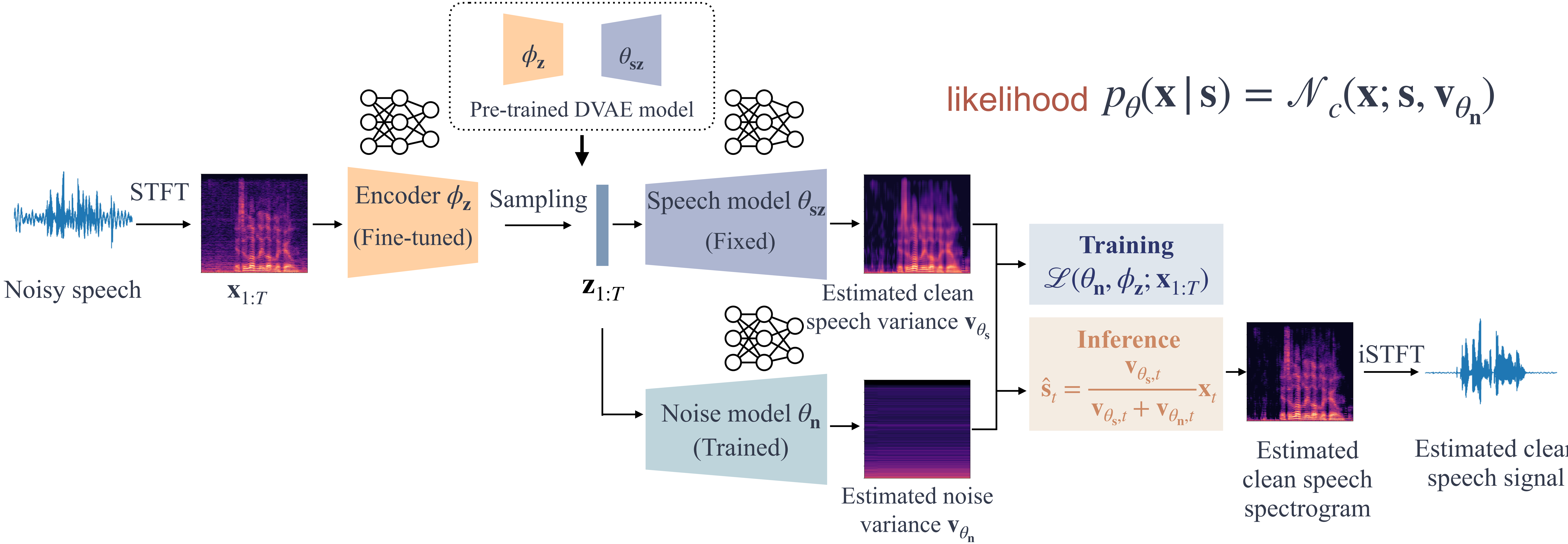
- **Pre-train a DVAE model on clean speech signals**



# Speech enhancement with Bayesian inference

- **Speech enhancement with the pre-trained DVAE and DDGM-based noise model**

likelihood $p_\theta(\mathbf{x} \mid \mathbf{s}) = \mathcal{N}_c(\mathbf{x}; \mathbf{s}, \mathbf{v}_{\theta_\mathbf{n}})$

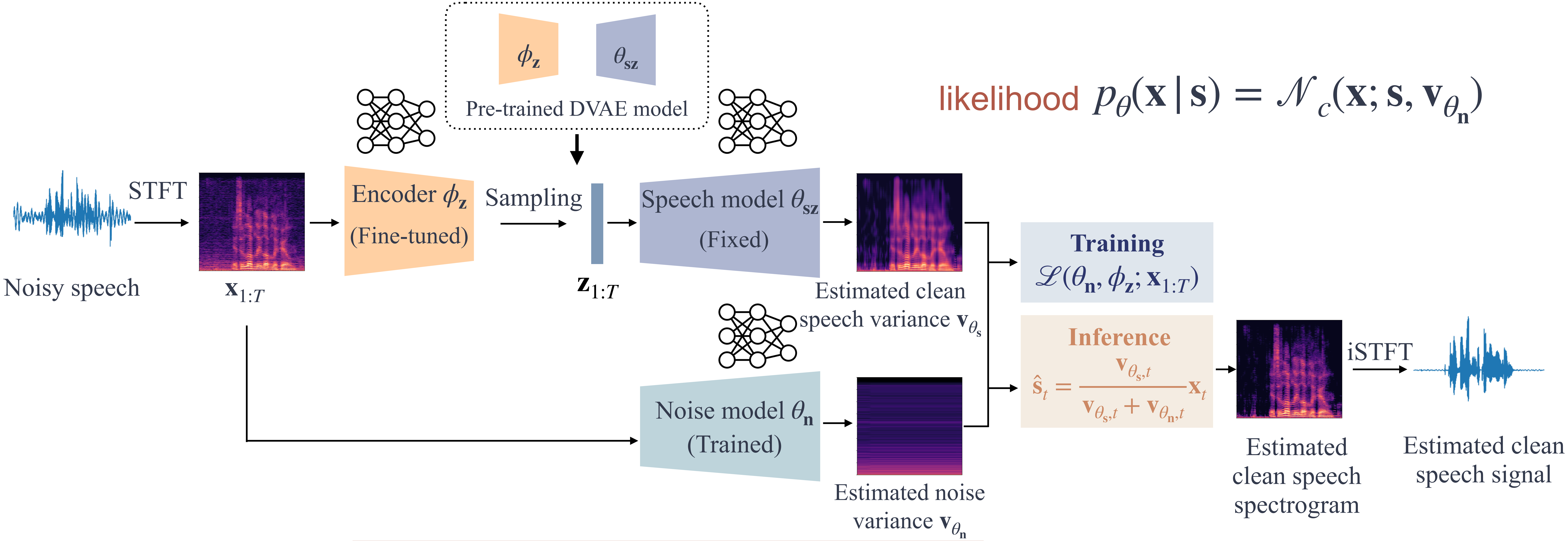



**DVAE latent variables (LV):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{z}_{1:T})$

# Speech enhancement with Bayesian inference

- **Speech enhancement with the pre-trained DVAE and DDGM-based noise model**

likelihood $p_\theta(\mathbf{x} \mid \mathbf{s}) = \mathcal{N}_c(\mathbf{x}; \mathbf{s}, \mathbf{v}_{\theta_\mathbf{n}})$

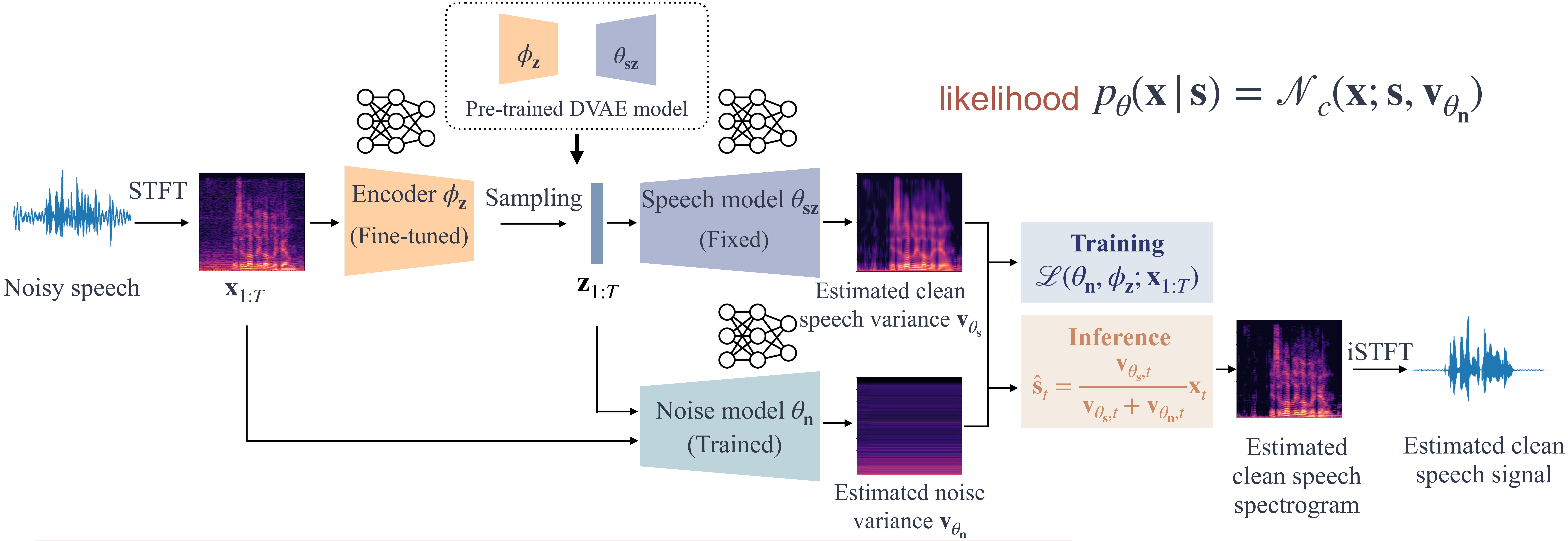



**Noisy observations (NO):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{x}_{1:t-1})$

# Speech enhancement with Bayesian inference

- **Speech enhancement with the pre-trained DVAE and DDGM-based noise model**

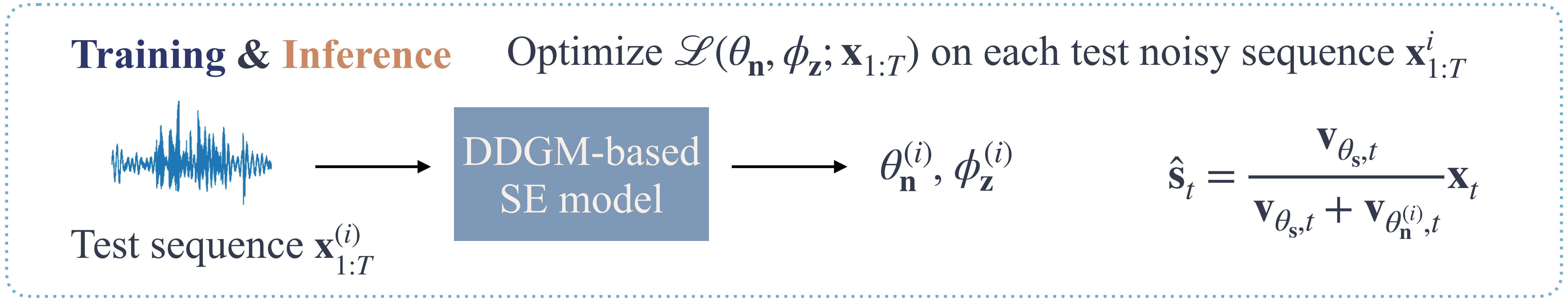
likelihood $p_\theta(\mathbf{x} \mid \mathbf{s}) = \mathcal{N}_c(\mathbf{x}; \mathbf{s}, \mathbf{v}_{\theta_\mathbf{n}})$



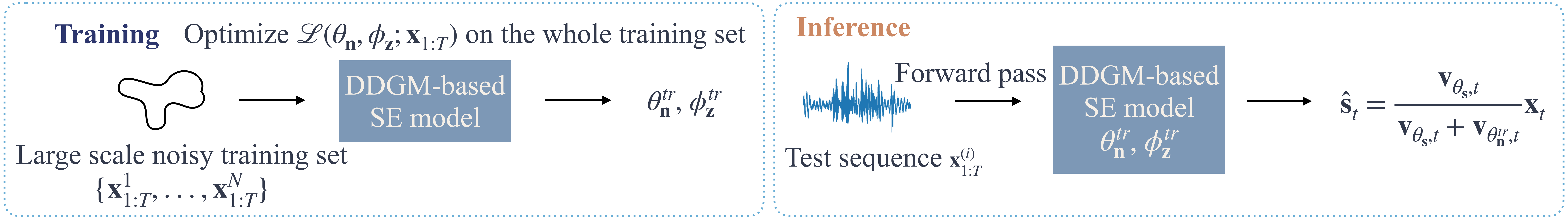

**Noisy observations and latent variables (NOLV):** $\mathbf{v}_{\theta_\mathbf{n},t} = \mathbf{v}_{\theta_\mathbf{n},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$

# Three training and evaluation configurations

- Unsupervised noise-agnostic (U-NA).




- Unsupervised noise-dependent (U-ND).




- U-NA fine-tuning after U-ND training (U-NDA).

# Experimental settings

## Datasets

- VoiceBank-DEMAND (VB-DMD)[48].
- WSJ0-QUT[49].

## Pre-processing

STFT coefficients: 64-ms sine window (1,024 samples) and 75%-overlap (256-sample shift).

## Baselines

- Supervised methods: Open-Unmix (UMX)[50] (LSTM-based method), MetricGAN+[51] (LSTM-based method), CDiffuSE[52] (diffusion-based method), SGMSE+[53] (diffusion-based method).
- Unsupervised methods: MetricGAN-U[54], NyTT[55], RVAE-VEM[56] (DVAE+NMF noise model).

## Evaluation metrics

- Enhancement performance: SI-SDR, PESQ (in [-0.5, 4.5]), extended short-time objective intelligibility(ESTOI)[57] (in [0, 1]).
- Computational efficiency: Real-time factor (RTF) which is the time required to process 1 second of audio.

[48] Cassia Valentini-Botinhao, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. *Proc. Speech Synthesis Workshop*. 2016.
[49] Simon Leglaive, et al. A recurrent variational autoencoder for speech enhancement. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2020.
[50] Fabian-Robert Stöter,et al. Open-Unmix – A reference implementation for music source separation. *J. Open Source Software*. 2019.
[51] Szu-Wei Fu, et al. MetricGAN+: An improved version of MetricGAN for speech enhancement. *Proc. Interspeech Conf.* 2021.
[52] Yen-Ju Lu, et al. Conditional diffusion probabilistic model for speech enhancement.*Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2022.
[53] Julius Richter, et al. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 2023.
[54] Szu-Wei Fu, et al. Unsupervised speech enhancement / dereverberation based only on noisy / reverberated speech. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2022.
[55] Takuya Fujimura, et al. A training strategy for DNN-based speech enhancement without clean speech. *Proc. Europ. Signal Process. Conf. (EUSIPCO)*. 2021
[56] Xiaoyu Bie, et al. Unsupervised speech enhancement using dynamical variational autoencoders. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 2022.
[57] Cees H. Taal, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.* 2011.

# Experimental results

## Comparison of different noise models with different training configuations

- **Different noise models**

RVAE-LV: $\mathbf{v}_{\theta_{\mathrm{n}},t} = \mathbf{v}_{\theta_{\mathrm{n}},t}(\mathbf{z}_{1:T})$

RVAE-NO: $\mathbf{v}_{\theta_{\mathrm{n}},t} = \mathbf{v}_{\theta_{\mathrm{n}},t}(\mathbf{x}_{1:t-1})$

RVAE-NOLV: $\mathbf{v}_{\theta_{\mathrm{n}},t} = \mathbf{v}_{\theta_{\mathrm{n}},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$

| Dataset | Training configuration | Model | SI-SDR ↑ | $\mathrm{PESQ}_{\mathrm{MOS}}$ ↑ | ESTOI ↑ |
|---|---|---|---|---|---|
| WSJ0-QUT | - | Noisy mixture | -2.6 | 1.83 | 0.50 |
| | U-NA | RVAE-LV | 5.4 | 2.31 | **0.65** |
| | | RVAE-NO | **6.0** | **2.33** | **0.65** |
| | | RVAE-NOLV | 5.5 | 2.31 | **0.65** |
| | U-ND | RVAE-LV | **5.3** | **2.25** | **0.60** |
| | | RVAE-NO | 3.7 | 2.11 | 0.58 |
| | | RVAE-NOLV | 4.9 | 2.11 | **0.60** |
| | U-NDA | RVAE-LV | **6.2** | **2.38** | 0.62 |
| | | RVAE-NO | 5.8 | 2.31 | **0.63** |
| | | RVAE-NOLV | **6.2** | 2.29 | 0.62 |
| VB-DMD | Noisy mixture | - | 8.4 | 3.02 | 0.79 |
| | U-NA | RVAE-LV | **17.5** | 3.23 | **0.82** |
| | | RVAE-NO | 17.3 | **3.25** | **0.82** |
| | | RVAE-NOLV | **17.5** | **3.25** | **0.82** |
| | U-ND | RVAE-LV | **17.4** | **3.24** | **0.81** |
| | | RVAE-NO | 16.7 | 3.03 | 0.79 |
| | | RVAE-NOLV | 16.9 | 3.04 | 0.79 |
| | U-NDA | RVAE-LV | **17.8** | **3.22** | **0.81** |
| | | RVAE-NO | 17.2 | 3.06 | 0.80 |
| | | RVAE-NOLV | 17.4 | 3.17 | **0.81** |

# Experimental results

## Comparison with baseline models

- **Different training configurations**

  Performance

  **U-NA** > **U-ND**

  Inference speed

  **U-NA** < < **U-ND**

  Further improvements

  **U-NDA**

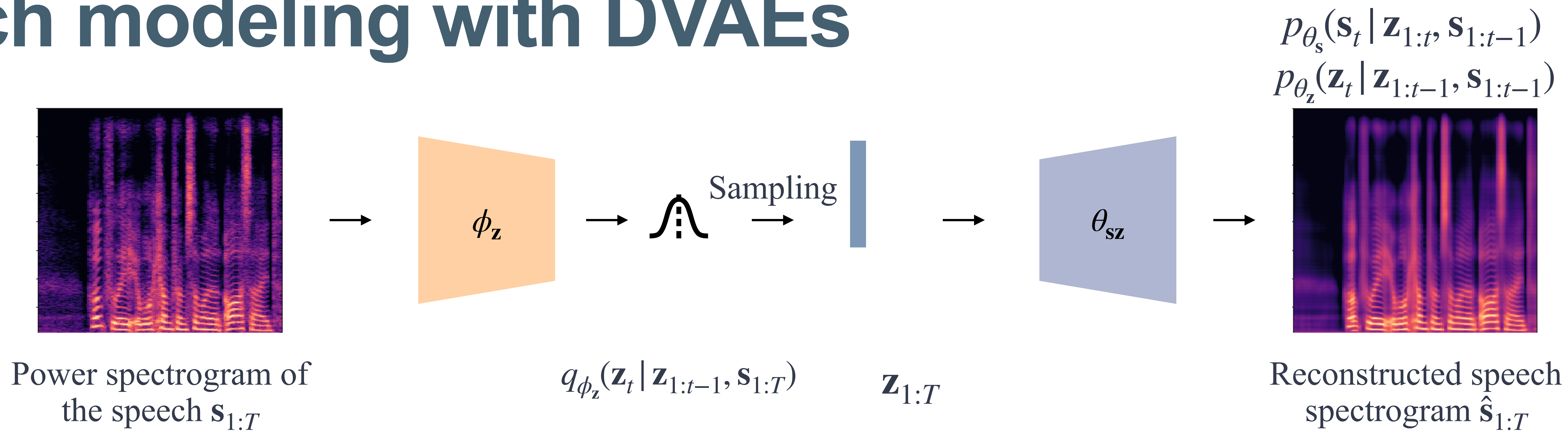
| Dataset | Model | Supervision | SI-SDR ↑ | $\text{PESQ}_{\text{MOS}}$ ↑ | ESTOI ↑ | # Iter. ↓ | RTF ↓ |
|---|---|---|---|---|---|---|---|
| WSJ0-QUT | Noisy mixture | - | -2.6 | 1.83 | 0.50 | - | - |
| | UMX | Supervised | 5.7 | 2.16 | 0.63 | | |
| | MetricGAN+ | Supervised | 3.6 | **2.83** | 0.60 | - | - |
| | RVAE-VEM | U-NA | 5.8 | 2.27 | 0.62 | 300 | 27.91 |
| | RVAE-LV | U-NA | 5.4 | 2.31 | **0.65** | 1000 | 89.42 |
| | | U-ND | 5.3 | 2.25 | 0.60 | **0** | **0.02** |
| | | U-NDA | **6.2** | 2.38 | 0.62 | 190 | 17.42 |
| VB-DMD | Noisy mixture | - | 8.4 | 3.02 | 0.79 | - | - |
| | UMX | Supervised | 14.0 | 3.18 | 0.83 | - | - |
| | MetricGAN+ | Supervised | 8.5 | **3.59** | 0.83 | - | - |
| | CDiffuSE | Supervised | 12.6 | - | 0.79 | - | - |
| | SGMSE+ | Supervised | 17.3 | - | **0.87** | - | 3.39 |
| | NyTT Xtra | U-ND | 17.7 | - | - | - | - |
| | MetricGAN-U | U-ND | 8.2 | 3.20 | 0.77 | - | - |
| | RVAE-VEM | U-NA | 17.1 | 3.23 | 0.81 | 100 | 9.55 |
| | RVAE-LV | U-NA | 17.5 | 3.23 | 0.82 | 900 | 81.62 |
| | | U-ND | 17.4 | 3.24 | 0.81 | **0** | **0.02** |
| | | U-NDA | **17.8** | 3.22 | 0.81 | 25 | 2.32 |

# Part 3

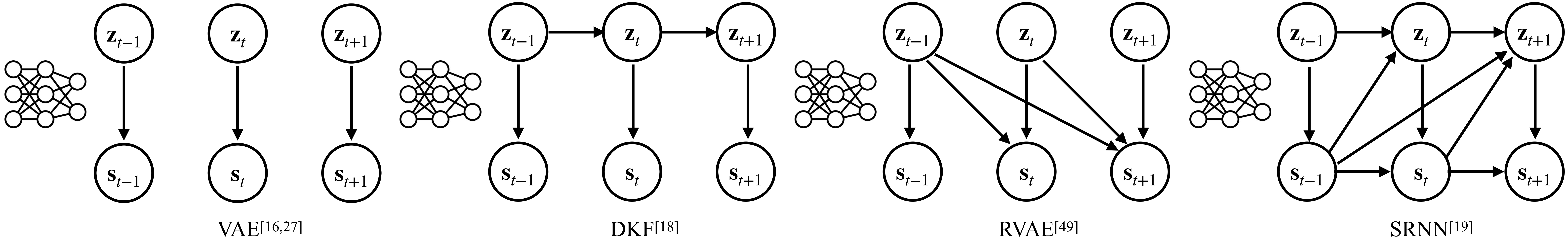
# Speech modeling with a hierarchical Transformer dynamical VAE

Xiaoyu Lin, Xiaoyu Bie, Simon Leglaive, Laurent Girin, and Xavier Alameda-Pineda. “Speech modeling with a hierarchical Transformer dynamical VAE.” In IEEE International Conference on Acoustics, Speech and Signal Processing, 2023.

# Speech modeling with DVAEs



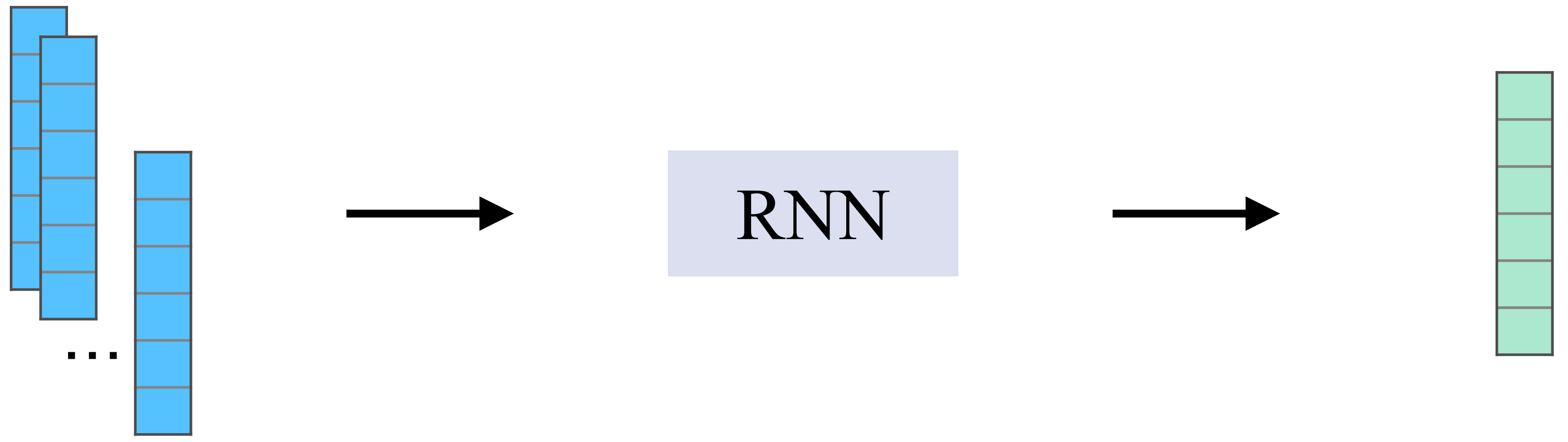

## Temporal dependencies of different DVAEs




[18] Rahul Krishnan, et al. Deep kalman filters. *Advances in Approx. Bayesian Infer*. 2015.
[19] Marco Fraccaro, et al. Sequential neural models with stochastic layers. *Advances in Neural Inform. Process. Systems (NeurIPS)*. 2016.
[16] Diederik P. Kingma, et al. Auto-encoding variational Bayes. *Proc. Int. Conf. Learn. Repres. (ICLR)*. 2014.
[27] Danilo Jimenez Rezende, et al. Stochastic backpropagation and approximate inference in deep generative models. *Proc. Int. Conf. Mach. Learn. (ICML)*. 2014.
[49] Simon Leglaive, et al. A recurrent variational autoencoder for speech enhancement. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2020.

# RNN-based auto-regressive (AR) model training issues

## Teacher-forcing (TF)[58] training procedure




Ground truth past values

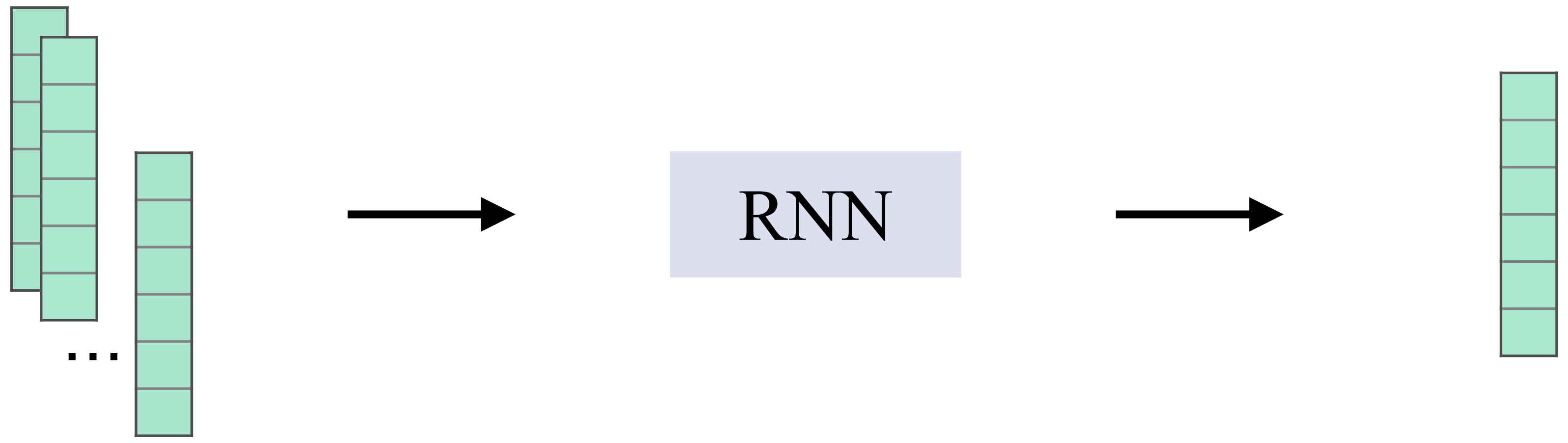
$\mathbf{s}_{1:t-1}$

$\hat{\mathbf{s}}_t = f(\mathbf{s}_{1:t-1})$

**Issue: At inference time we can only use the generated previous values to predict $\hat{\mathbf{s}}_t$, which will cause large accumulated errors.**

## Scheduled-sampling (SS)[59] training procedure

Gradually replace the GT past values by predicted past values along training iterations.




Predicted past values

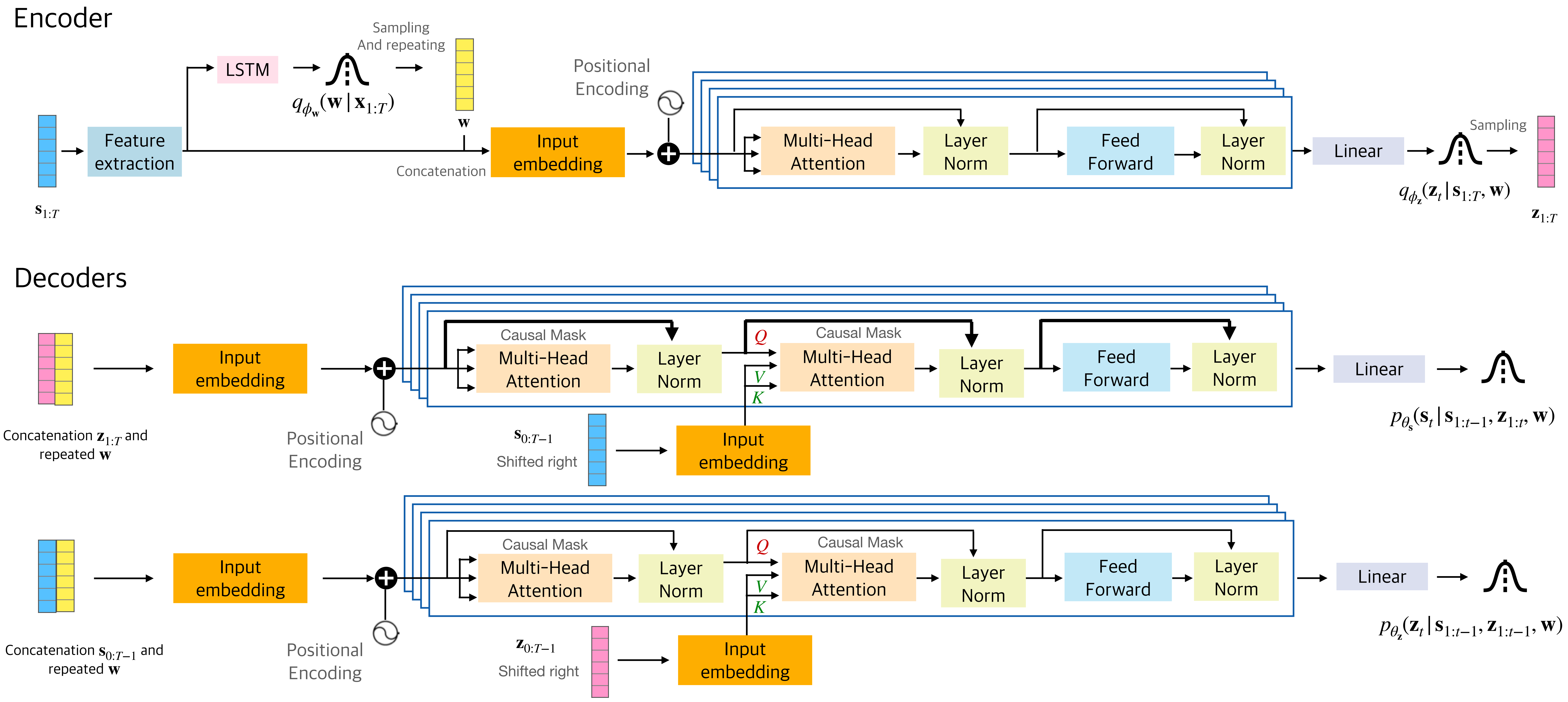
$\hat{\mathbf{s}}_{1:t-1}$

$\hat{\mathbf{s}}_t = f(\hat{\mathbf{s}}_{1:t-1})$

**Limitations: requirements of a well-designed sampling scheduler to guarantee the performance.**

[58] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comp.* 1989.
[59] Samy Bengio, et al. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Inform. Process. Systems (NeurIPS)*. 2015.

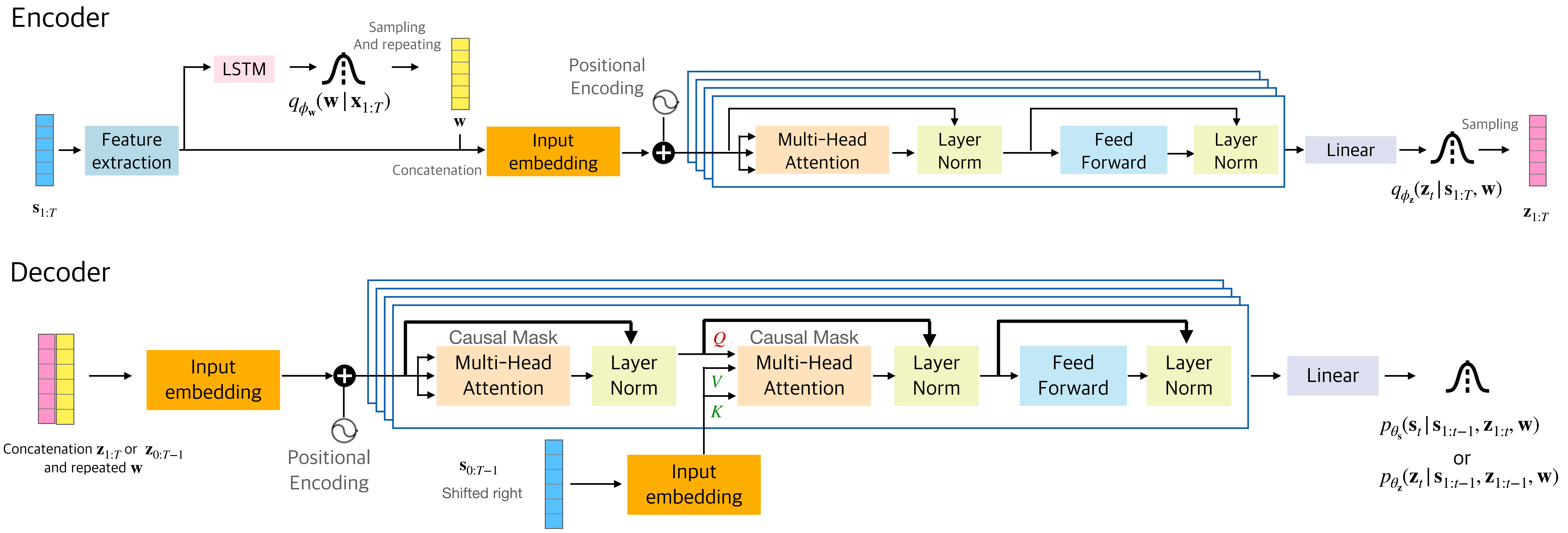
# HiT-DVAE model[60]






[60] Xiaoyu Bie, et al. HiT-DVAE: Human motion generation via Hierarchical Transformer Dynamical VAE. *arXiv preprint arxiv:2204.01565*, 2022.

# LigHT-DVAE model




## Model training by maximizing the Evidence Lower BOund (ELBO)

$$\mathscr{L}(\theta,\phi;\mathbf{s}_{1:T}) = -D_{\mathrm{KL}}(q_{\phi_\mathbf{w}}(\mathbf{w}\,|\,\mathbf{s}_{1:T})p_{\theta_\mathbf{w}}(\mathbf{w})) - \sum_{t=1}^{T} \mathbb{E}_{q_{\phi_\mathbf{z}}q_{\phi_\mathbf{w}}}\Big[d_{\mathrm{IS}}(|\mathbf{s}_t|^2, \mathbf{v}_{\theta_\mathrm{s},t}) + D_{\mathrm{KL}}(q_{\phi_\mathbf{z}}(\mathbf{z}_t\,|\,\mathbf{s}_{1:T},\mathbf{w}) \parallel p_{\theta_\mathbf{z}}(\mathbf{z}_t\,|\,\mathbf{s}_{1:t-1},\mathbf{z}_{1:t-1},\mathbf{w}))\Big]$$

regularization term for $\mathbf{w}$ | reconstruction term | regularization term for $\mathbf{z}$

# Experimental settings

## Datasets

- Wall Street Journal (WSJ0) dataset.
- Voice Bank (VB) corpus[61].

## Baselines

VAE, DKF, RVAE, SRNN (trained in SS), SRNN (trained in TF).

## Evaluation metrics

- Speech analysis-resynthesis: RMSE, SI-SDR, PESQ, ESTOI.
- Speech generation: Fréchet Deep Speech Distance (FDSD)[62].

[61] Christophe Veaux, et al. The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database. *Proceedings of International Committee for Co-ordination and Standardisation of Speech Databases*, 2013.
[62] Mikołaj Bińkowski, et al. High fidelity speech synthesis with adversarial networks. *Proc. Int. Conf. Learn. Repres. (ICLR).* 2020

# Experimental results for speech analysis-resynthesis

| Dataset | Model | RMSE ↓ | SI-SDR ↑ | PESQ ↑ | ESTOI ↑ |
|---|---|---|---|---|---|
| WSJ0 | VAE | 0.040 | 7.4 | 3.28 | 0.88 |
| | DKF | 0.037 | 8.3 | 3.51 | **0.91** |
| | RVAE | 0.034 | 8.9 | 3.53 | **0.91** |
| | SRNN (SS) | 0.036 | 8.7 | **3.57** | **0.91** |
| | SRNN (TF) | 0.061 | 2.6 | 2.53 | 0.76 |
| | HiT-DVAE (TF) | 0.031 | 10.0 | 3.52 | **0.91** |
| | LigHT-DVAE (TF) | **0.030** | **10.1** | 3.55 | **0.91** |
| VB | VAE | 0.052 | 8.4 | 3.24 | 0.89 |
| | DKF | 0.048 | 9.3 | 3.44 | 0.91 |
| | RVAE | 0.050 | 8.9 | 3.39 | 0.90 |
| | SRNN (SS) | 0.044 | 10.1 | 3.42 | 0.91 |
| | SRNN (TF) | 0.102 | -0.1 | 2.15 | 0.75 |
| | HiT-DVAE (TF) | 0.039 | 11.4 | **3.60** | **0.93** |
| | LigHT-DVAE (TF) | **0.038** | **11.6** | 3.58 | **0.93** |

# Experimental results for speech generation

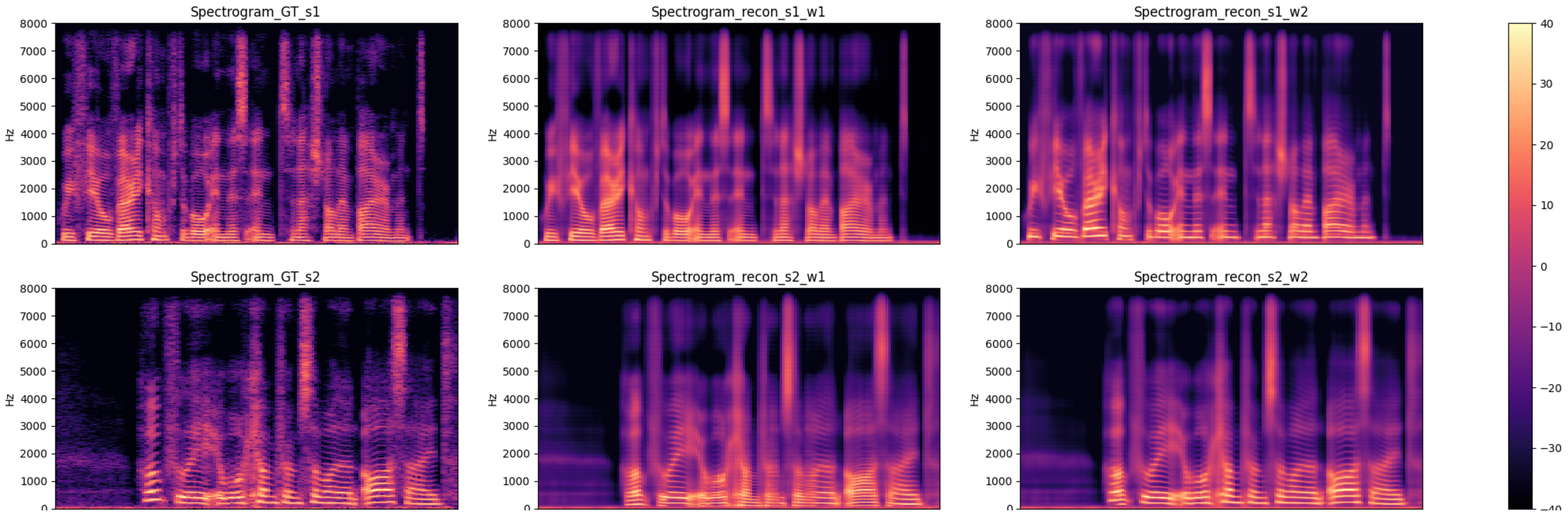
| Model | FDSD ↓ |
|---|---|
| VAE | 70.92 ± 0.44 |
| DKF | 32.78 ± 0.28 |
| RVAE | 45.75 ± 0.11 |
| SRNN (SS) | 25.28 ± 0.19 |
| SRNN (TF) | 25.53 ± 0.13 |
| HiT-DVAE | **22.50 ± 0.26** |
| LigHT-DVAE | 29.22 ± 0.26 |
| VB Test (exact phase) | 4.11 ± 0.14 |
| VB Test (Griffin-Lim) | 4.11 ± 0.15 |

Power spectrograms generated by the models and phase reconstructed with the Griffin-Lim[63] algorithm.



[63] Daniel W. Griffin and Jae S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.* 1984.

# Investigation on the role of $\mathbf{w}$




Swap the $\mathbf{w}$ to reconstruct the spectrograms.

# 04.
# Conclusion and Discussions

# A learning framework based on Bayesian inference

- Model $p_\theta(\mathbf{o} \mid \mathbf{s})$ with domain specific knowledge.

likelihood

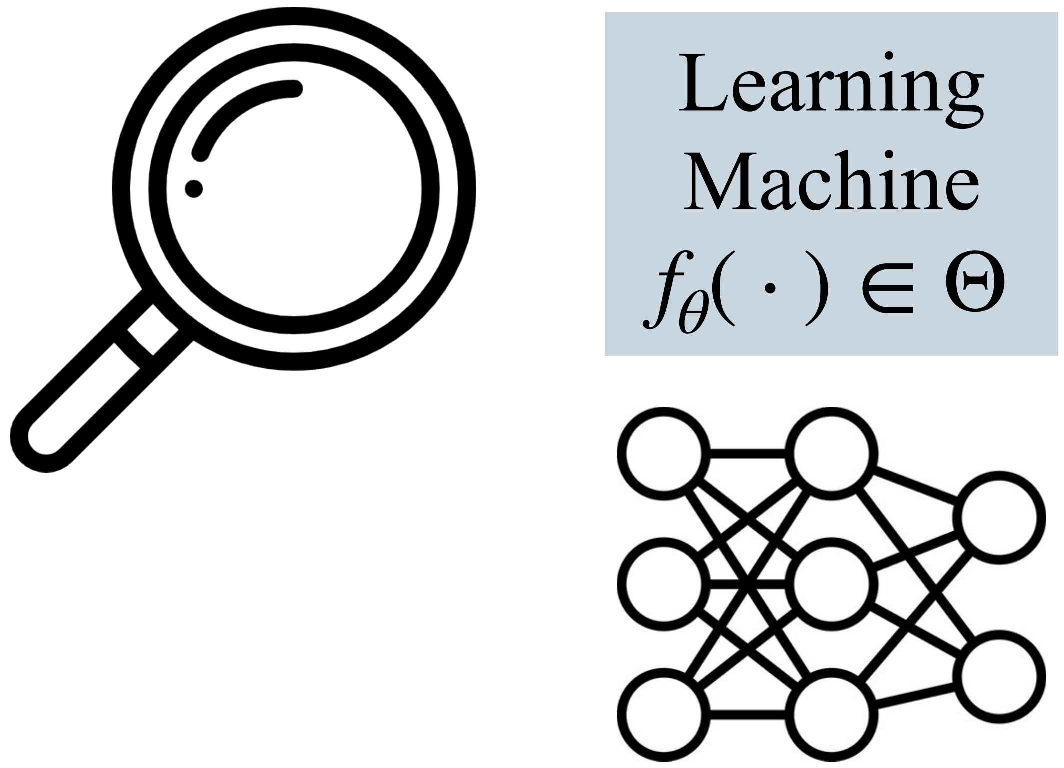
prior

$$p_\theta(\mathbf{s} \mid \mathbf{o}) = \frac{p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s}) ds}$$

posterior

marginal likelihood / evidence

# Interpretability




=

?

**Interpretable AI**

Bayesian inference methods are inherently interpretable.

# A learning framework based on Bayesian inference

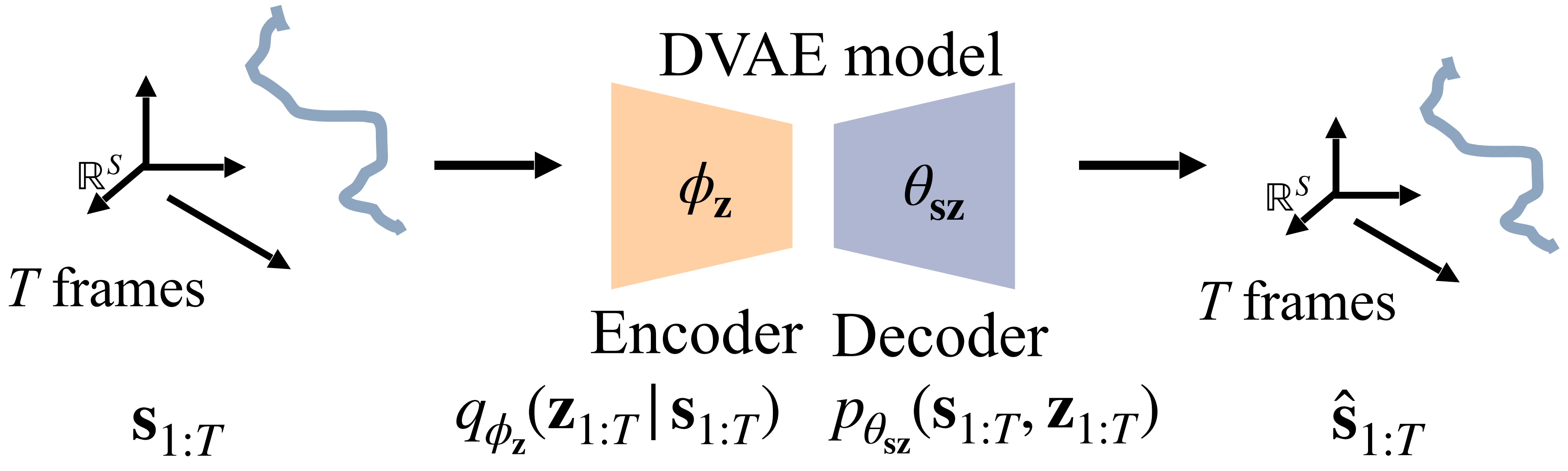
likelihood prior

$$p_\theta(\mathbf{s}\,|\,\mathbf{o}) = \frac{p_\theta(\mathbf{o}\,|\,\mathbf{s})p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{o}\,|\,\mathbf{s})p_\theta(\mathbf{s})d\mathbf{s}}$$

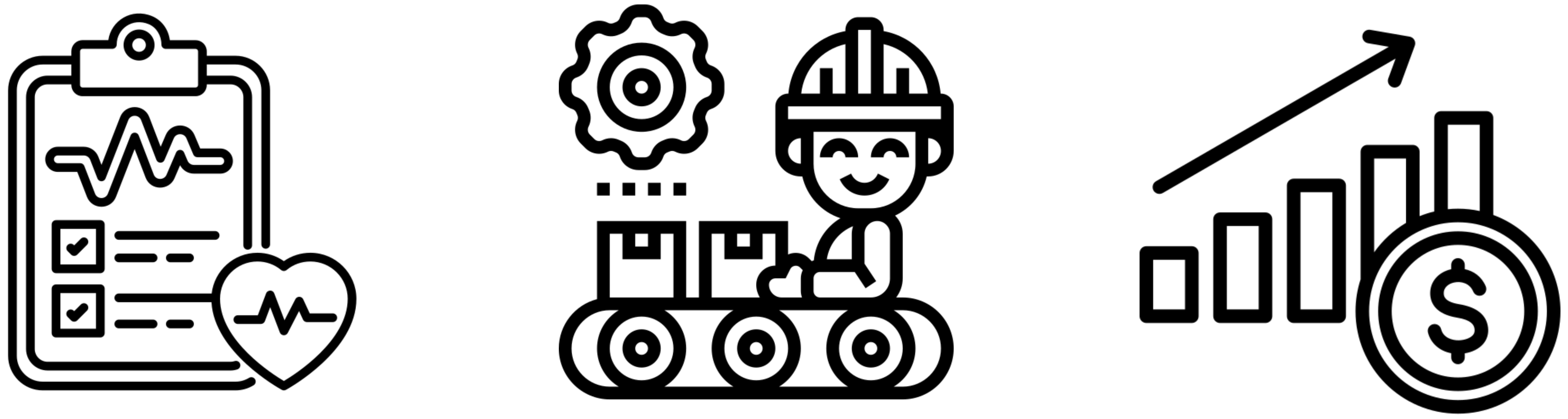
posterior

marginal likelihood / evidence

- Model $p_\theta(\mathbf{o}\,|\,\mathbf{s})$ with domain specific knowledge.
- Model $p_\theta(\mathbf{s})$ with a dynamical variational auto-encoder (DVAE).



# Data efficiency



**Health care** **Industrial production** **Finance**

?

Un-/weakly supervised learning framework.
No requirement for very large annotated training dataset.

# A learning framework based on Bayesian inference

likelihood

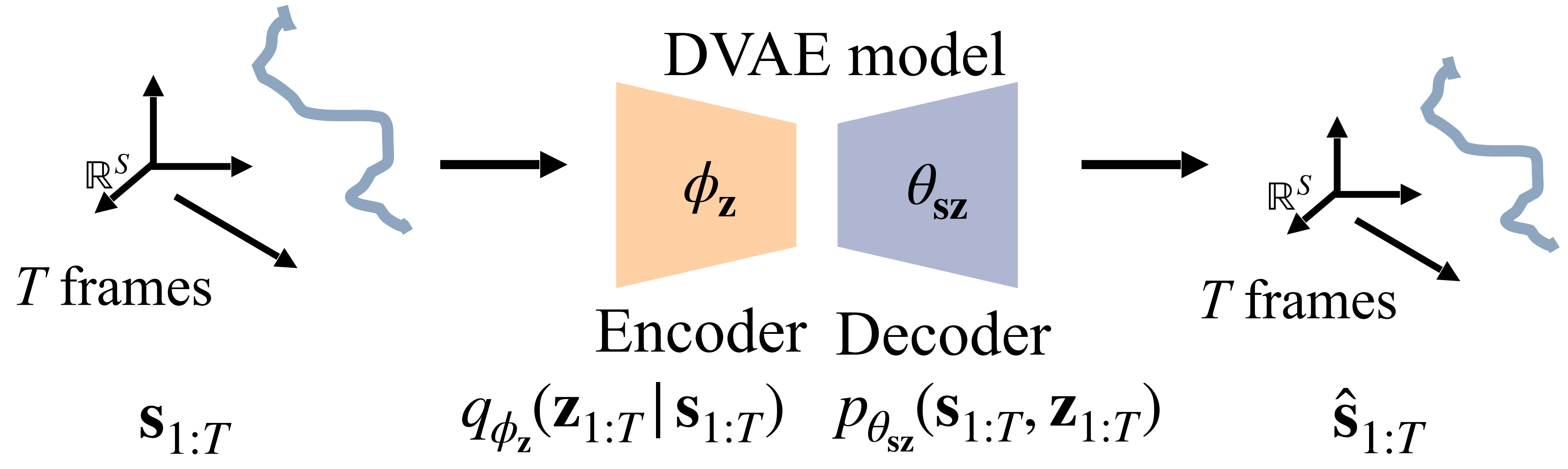
prior

$$p_\theta(\mathbf{s} \,|\, \mathbf{o}) = \frac{p_\theta(\mathbf{o} \,|\, \mathbf{s}) p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{o} \,|\, \mathbf{s}) p_\theta(\mathbf{s}) ds}$$

posterior

marginal likelihood / evidence

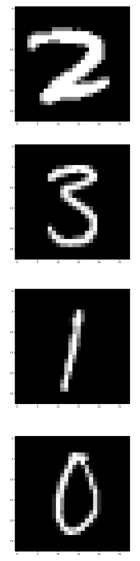
- Model $p_\theta(\mathbf{o} \,|\, \mathbf{s})$ with domain specific knowledge.
- Model $p_\theta(\mathbf{s})$ with a dynamical variational auto-encoder (DVAE).




- Infer $p_\theta(\mathbf{s} \,|\, \mathbf{o})$ with variational inference methodology
  - VEM for MOT and SC-ASS
  - Gradient-based optimization for SE

# Out-of-distribution generalization

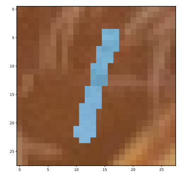
**Training**



i.i.d. training data samples

$(\mathbf{x}_1^{train}, \ldots, \mathbf{x}_N^{train}) \sim p(\mathbf{x})$

**Inference**



test data samples $\mathbf{x}^{test} \sim q(\mathbf{x})$

Learning Machine

$f_\theta(\cdot) \in \Theta$

$\hat{\mathbf{y}}^{train} = f_\theta(\mathbf{x}^{train})$

$\hat{\mathbf{y}}^{test} = f_\theta(\mathbf{x}^{test})$

?

**Out-of-distribution Generalization**

Integrating the pre-trained DVAE model into another LVGM has some link to the out-of-distribution generalization problem.

# A learning framework based on Bayesian inference

likelihood

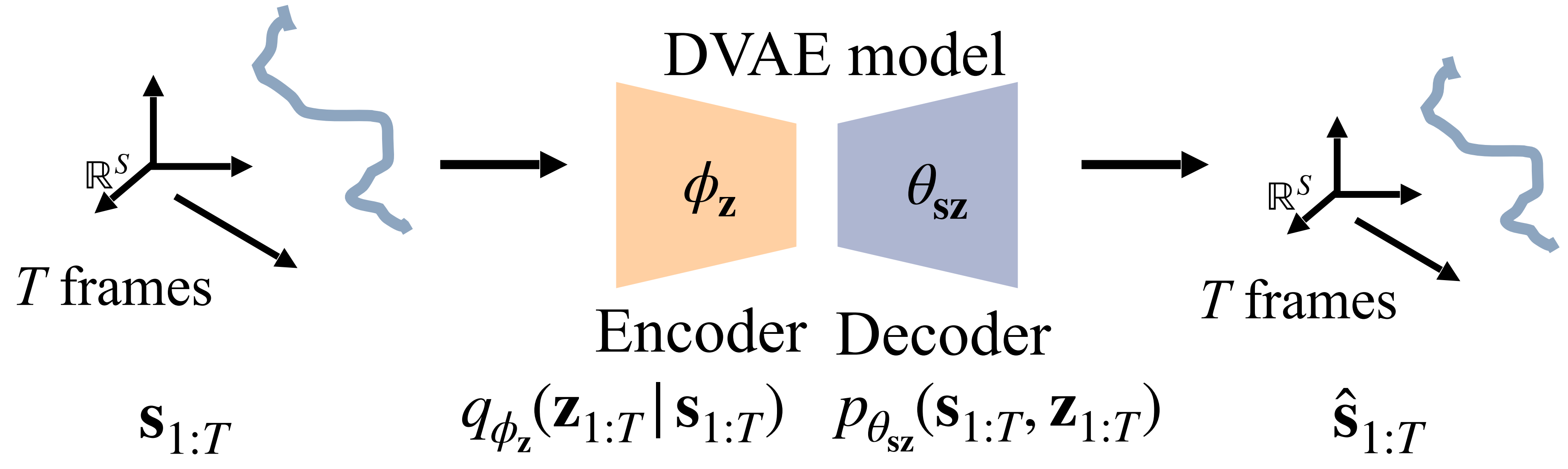
prior

$$p_\theta(\mathbf{s} \mid \mathbf{o}) = \frac{p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s})}{\int p_\theta(\mathbf{o} \mid \mathbf{s}) p_\theta(\mathbf{s}) ds}$$

posterior

marginal likelihood / evidence

- Model $p_\theta(\mathbf{o} \mid \mathbf{s})$ with domain specific knowledge.
- Model $p_\theta(\mathbf{s})$ with a dynamical variational auto-encoder (DVAE).




- Infer $p_\theta(\mathbf{s} \mid \mathbf{o})$ with variational inference methodology
  - VEM for MOT and SC-ASS.
  - Gradient-based optimization for SE.
- A novel DVAE architecture combined with Transformers: HiT/LigHT-DVAE.

# Advantages and limitations of this method

## Advantages

- **Data-frugal**: no need for large amount of annotated data.
- **Interpretability**: the possibility of incorporating human-level prior knowledge into the model.

## Limitations

- **Computational complexity**: the VEM algorithm can be very time consuming.
- **Subpar performance** compared to fully-supervised methods.

## Remarks

- The model's performance highly depends on the robustness of the pre-trained DVAE models.
- The latent variables learned by the DVAE models are still not well understood[64, 65, 66].

[64] Irina Higgins, et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. *Proc. Int. Conf. Learn. Repres. (ICLR)*. 2017.
[65] Shengjia Zhao, et al. Infovae: Balancing learning and inference in variational autoencoders.*Proc. AAAI Conf. Artif. Intell.* 2019.
[66] Yixin Wang, et al. Posterior Collapse and Latent Variable Non-identifiability. *Advances in Neural Inform. Process. Systems (NeurIPS)*. 2021

# 05.
# Future Research Direction

# Some reflections on the future research directions

- What are the other learning principles / paradigms that can generalize well for out-of-distribution data samples (strong generalization ability)[67,68,69,70]?
- How to better understand the latent representations learned by the DVAE models and other generative models[71,72]?
- What are the potential pathways to make the AI systems more robust, reliable and controllable so that they can be applied to more risk-sensitive domains[73,74]?

[67] Judea Pearl. Causal inference in statistics: An overview. 2009.
[68] Yishay Mansour, et al. Domain adaptation: Learning bounds and algorithms. *Proc. Conf. Learn. Theory (COLT)*. 2009.
[69] Martin Arjovsky, et al. Invariant risk minimization. *arXiv preprint arXiv:1907.02893.* 2019.
[70] Peng Cui, et al. Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell*. 2022.
[71] Ilyes Khemakhem, et al. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *Proc. Int. Conf. Mach. Learn. (ICML)*. 2020.
[72] Thibaut Issenhuth, et al. Unveiling the Latent Space Geometry of Push-Forward Generative Models. *Proc. Int. Conf. Mach. Learn. (ICML)*. 2023.
[73] Aleksander Madry, et al. Towards Deep Learning Models Resistant to Adversarial Attacks. *Proc. Int. Conf. Learn. Repres. (ICLR)*. 2018.
[74] Gregory Falco, et al. Governing AI safety through independent audits. *Nat. Mach. Intell.* 2021.

# Thanks for your attention.