Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation

Xiaoyu Lin, Laurent Girin, Xavier Alameda-Pineda
INRIA, Univ. Grenoble-Alpes

February 27th, 2024

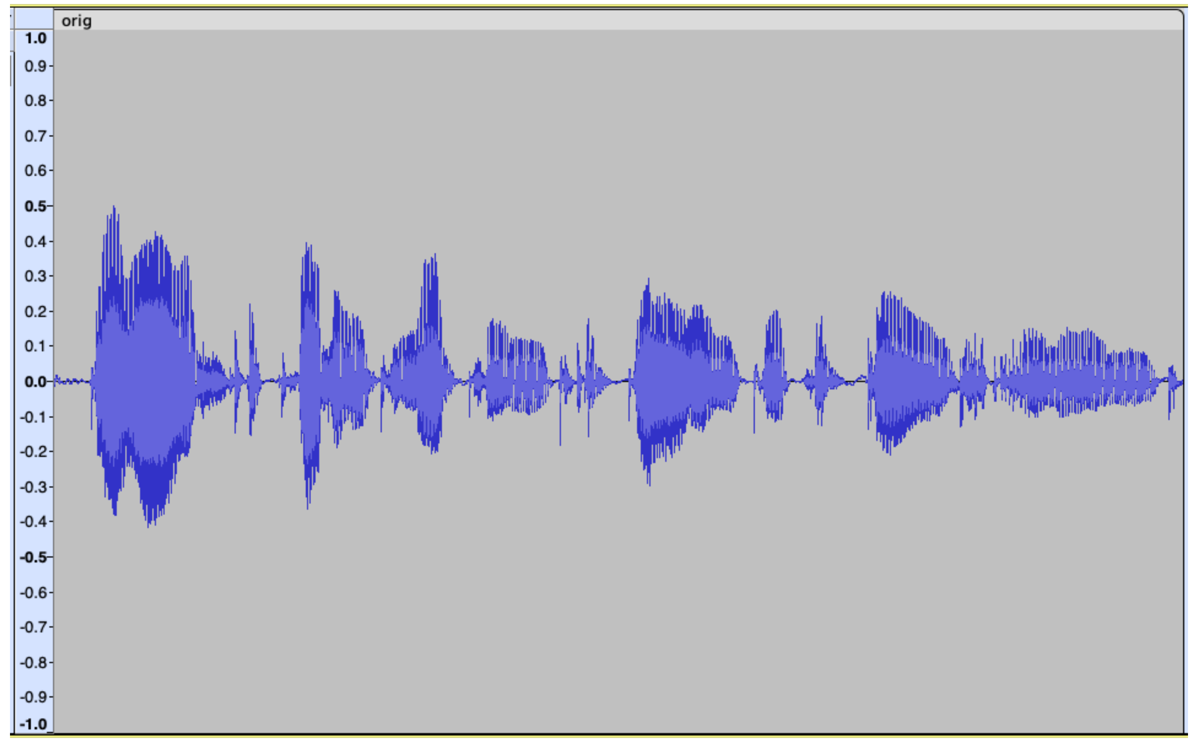
[1] Lin, X., Girin, L., & Alameda-Pineda, X., 2023. Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation. *Transactions on Machine Learning Research*.

Probabilistic Generative Models

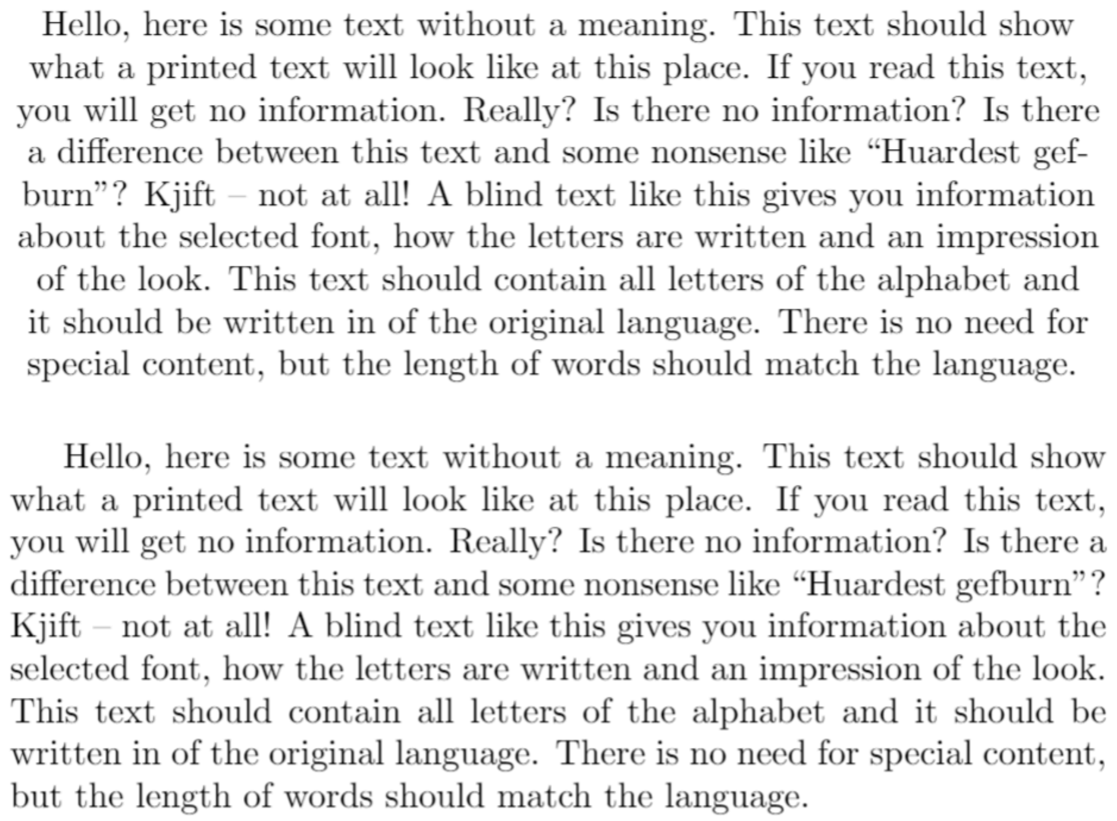
Motivations

- Understand complex real-world data

Image



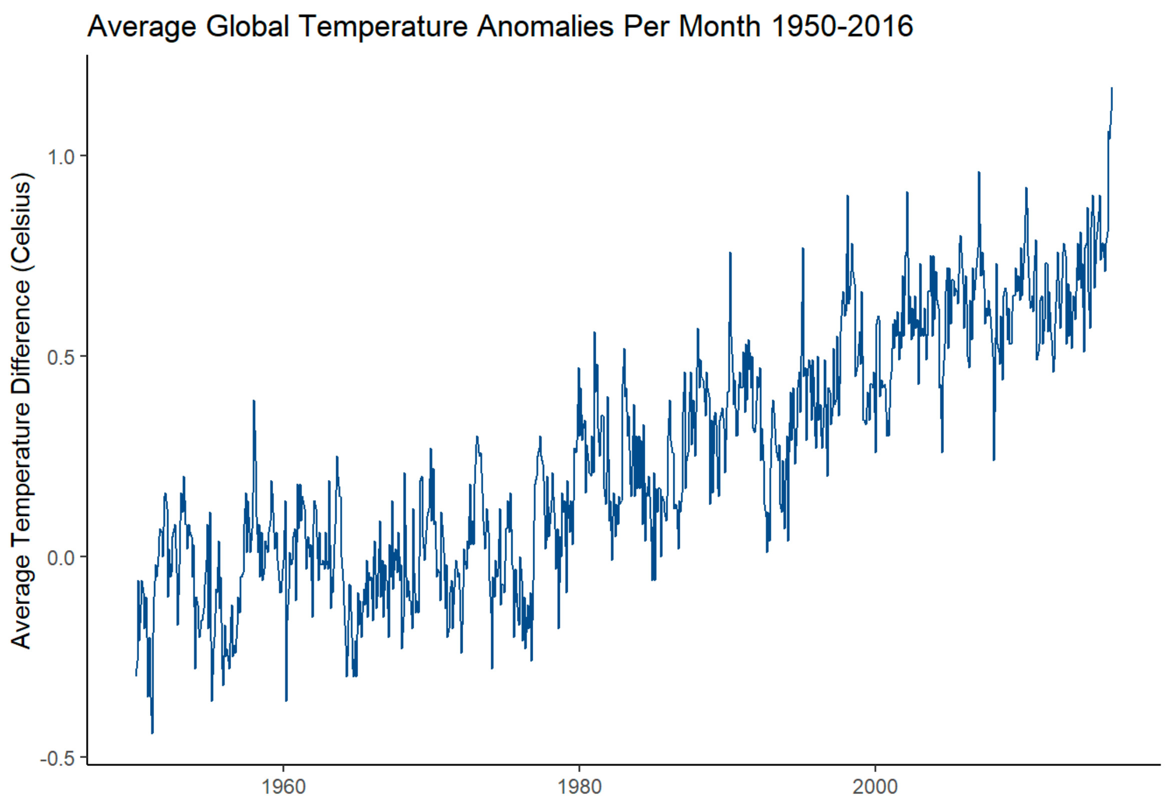
Audio



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

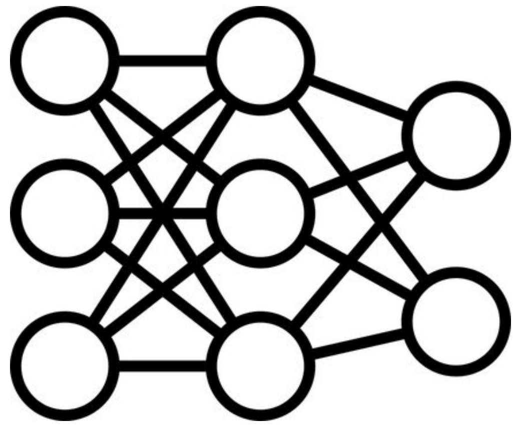


Time series

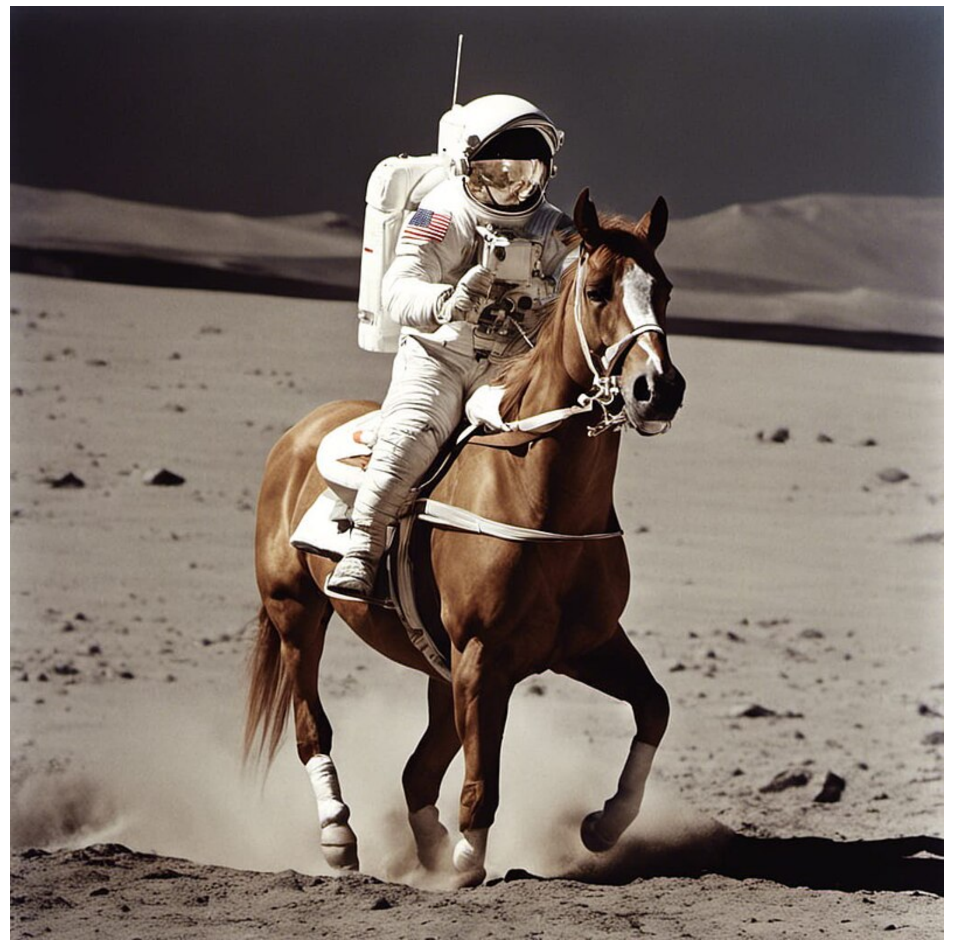
Motivations

- Understand complex real-world data
- Generate new data points

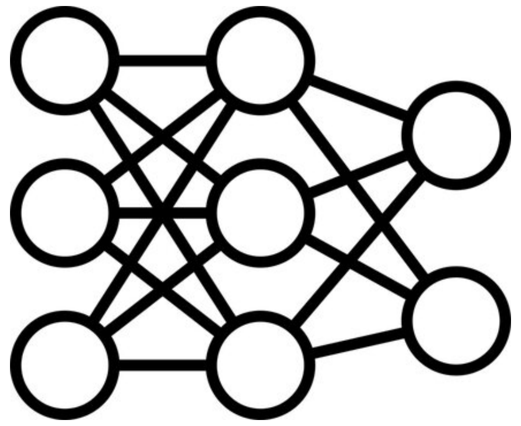
“An astronaut riding a horse”



generative model



“An 80s driving pop song with heavy drums and synth pads in the background”

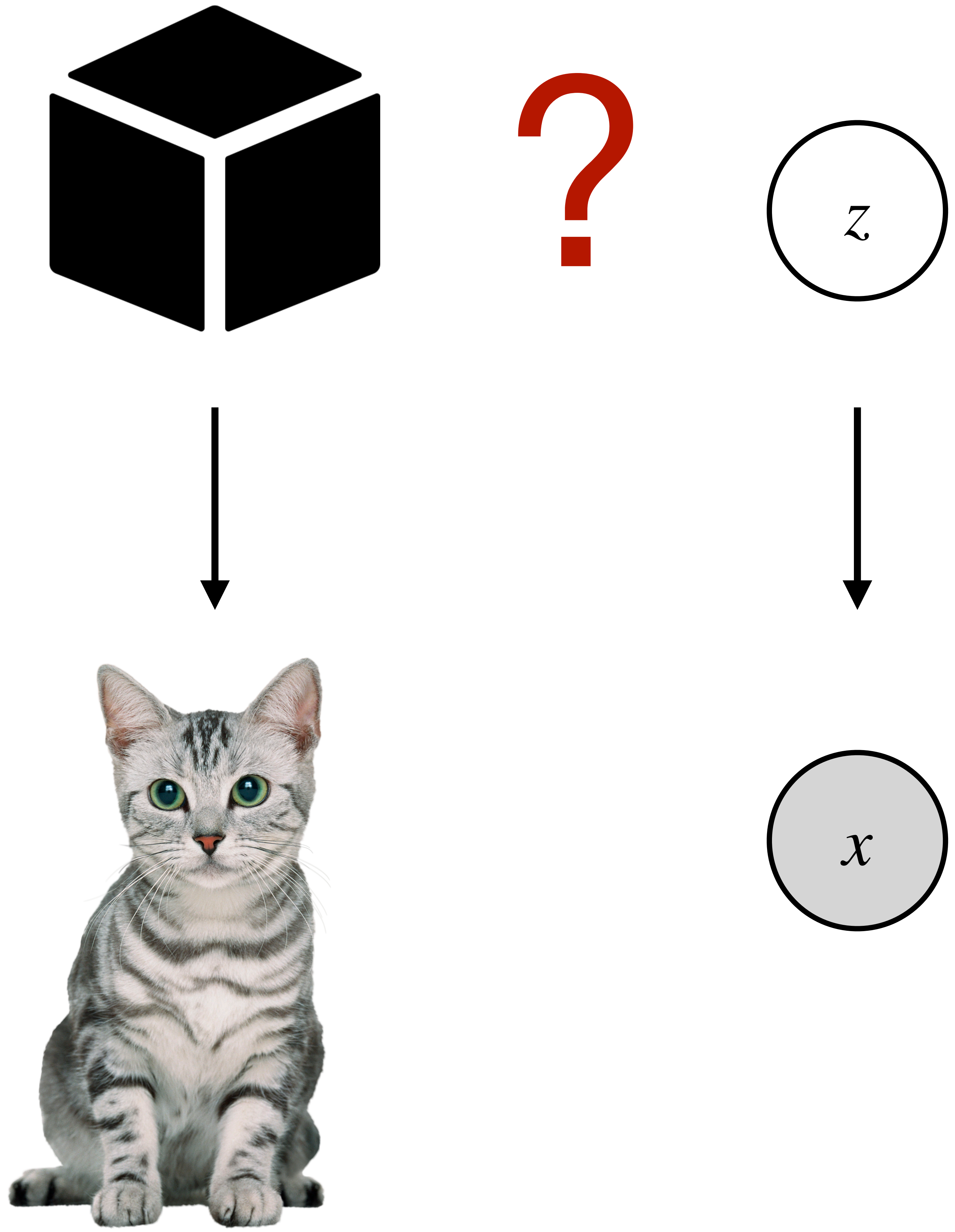


generative model

* Examples from DALLE 2 and MusicGen

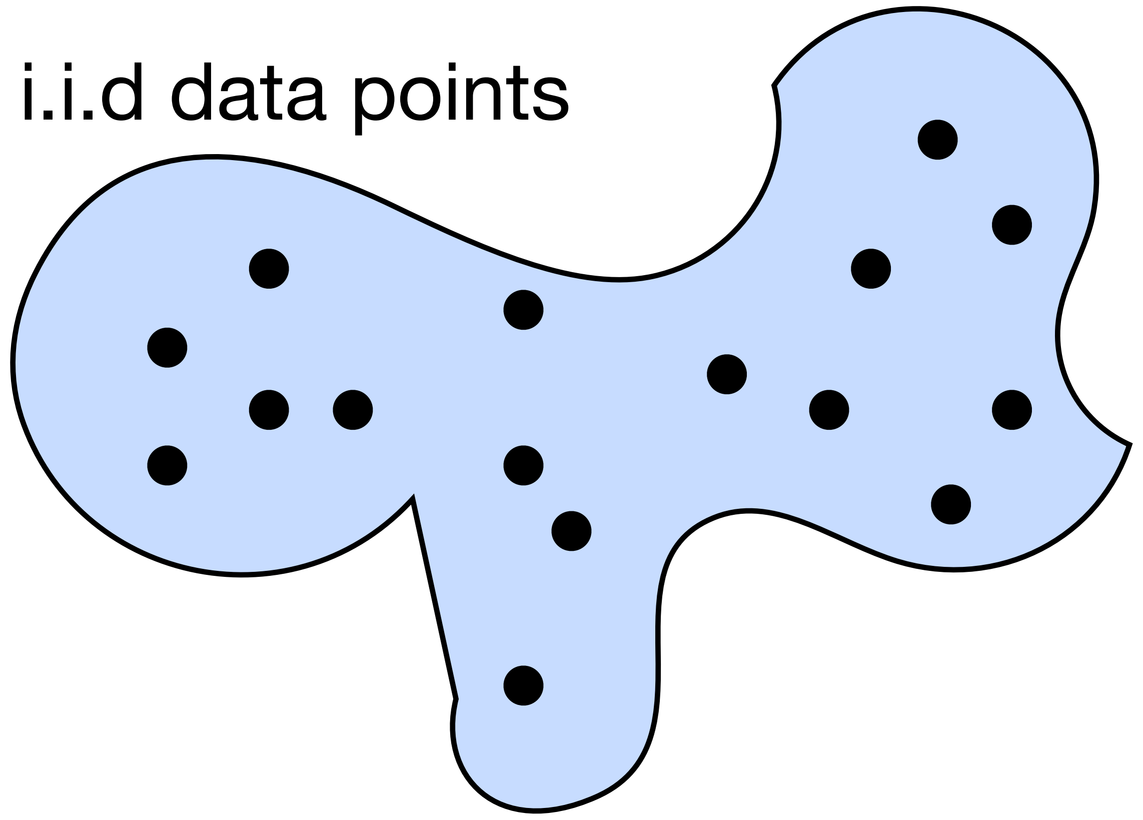
Motivations

- Understand complex real-world data
- Generate new data points
- Discover unknown quantities / data representations



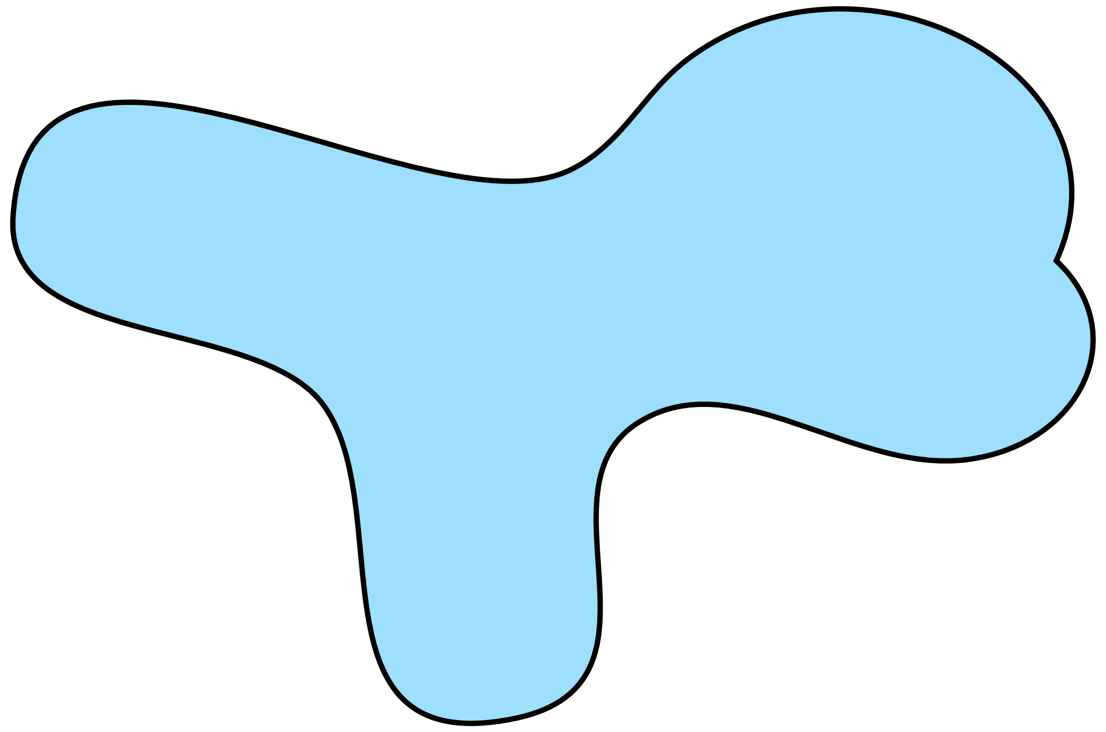
Approaches

- Implicit generative models
  - Generative Adversarial Networks (GANs)
- Explicit generative models: explicitly model the probability density function (PDF)



True data distribution

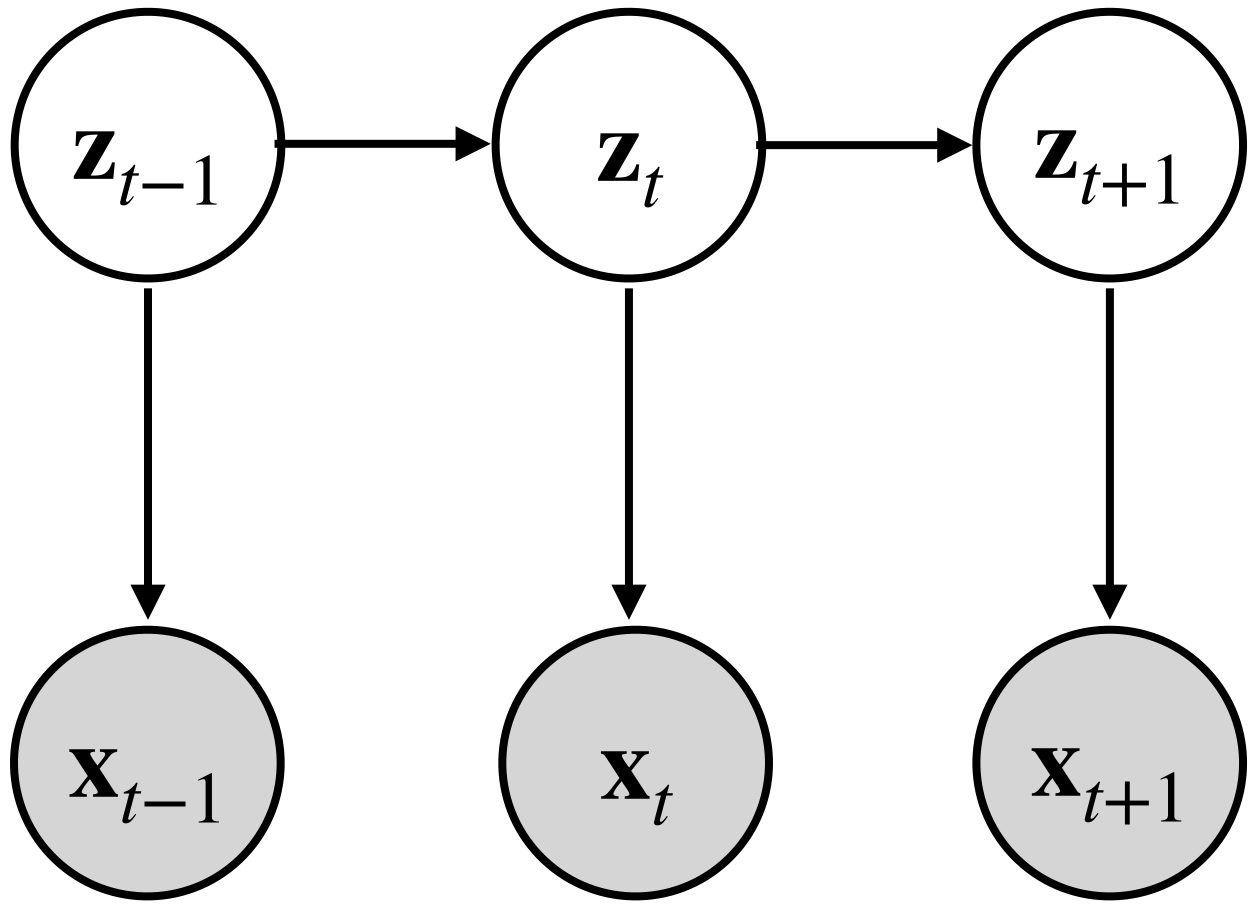
$p_{data}(\mathbf{x})$



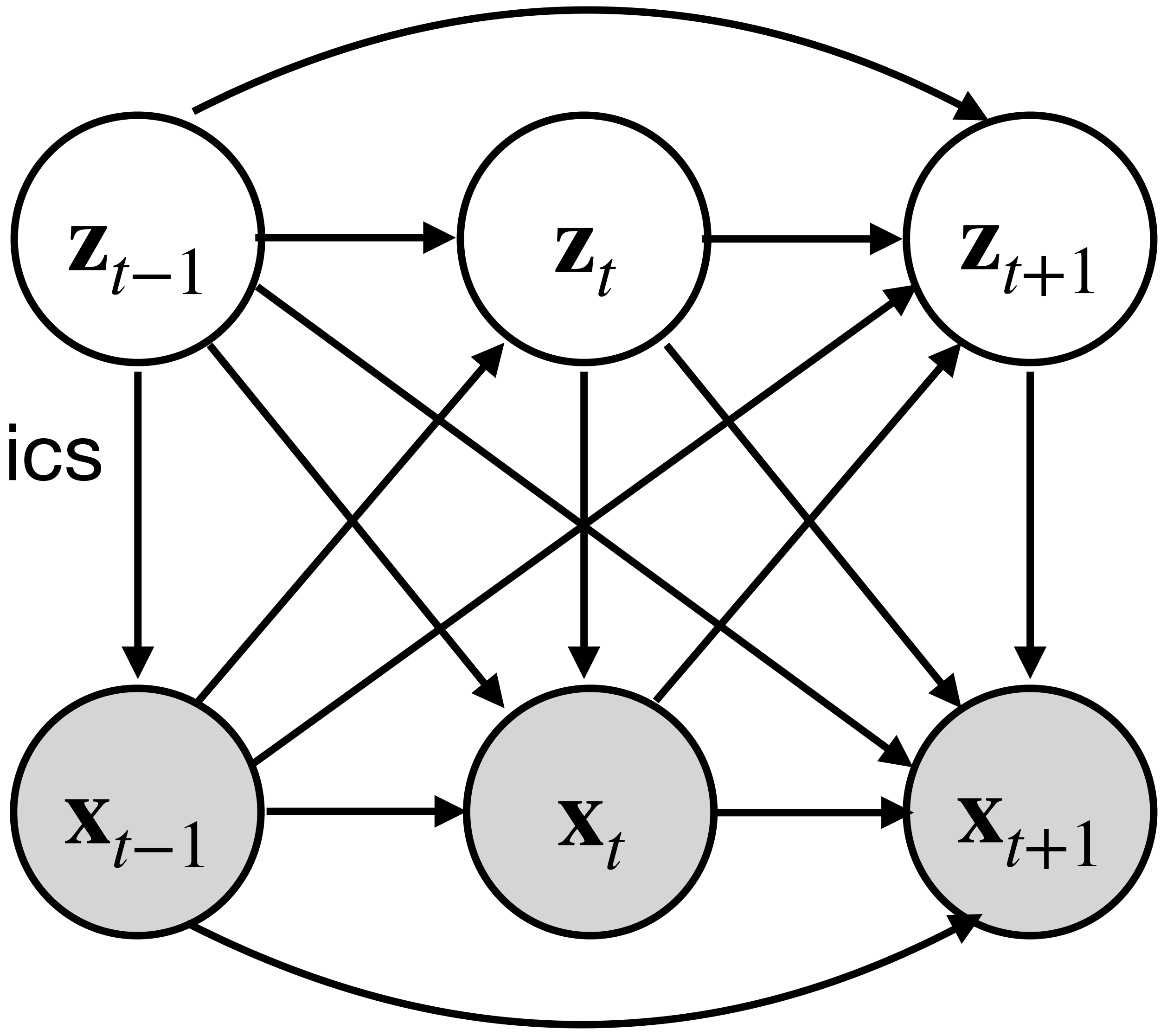
Parametric probabilistic model

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

Example: probabilistic modeling of sequential data



Non-linear dynamics



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

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

State Space Models
(SSM)

$\mathbf{z}$ discrete

Hidden Markov Model
(HMM)

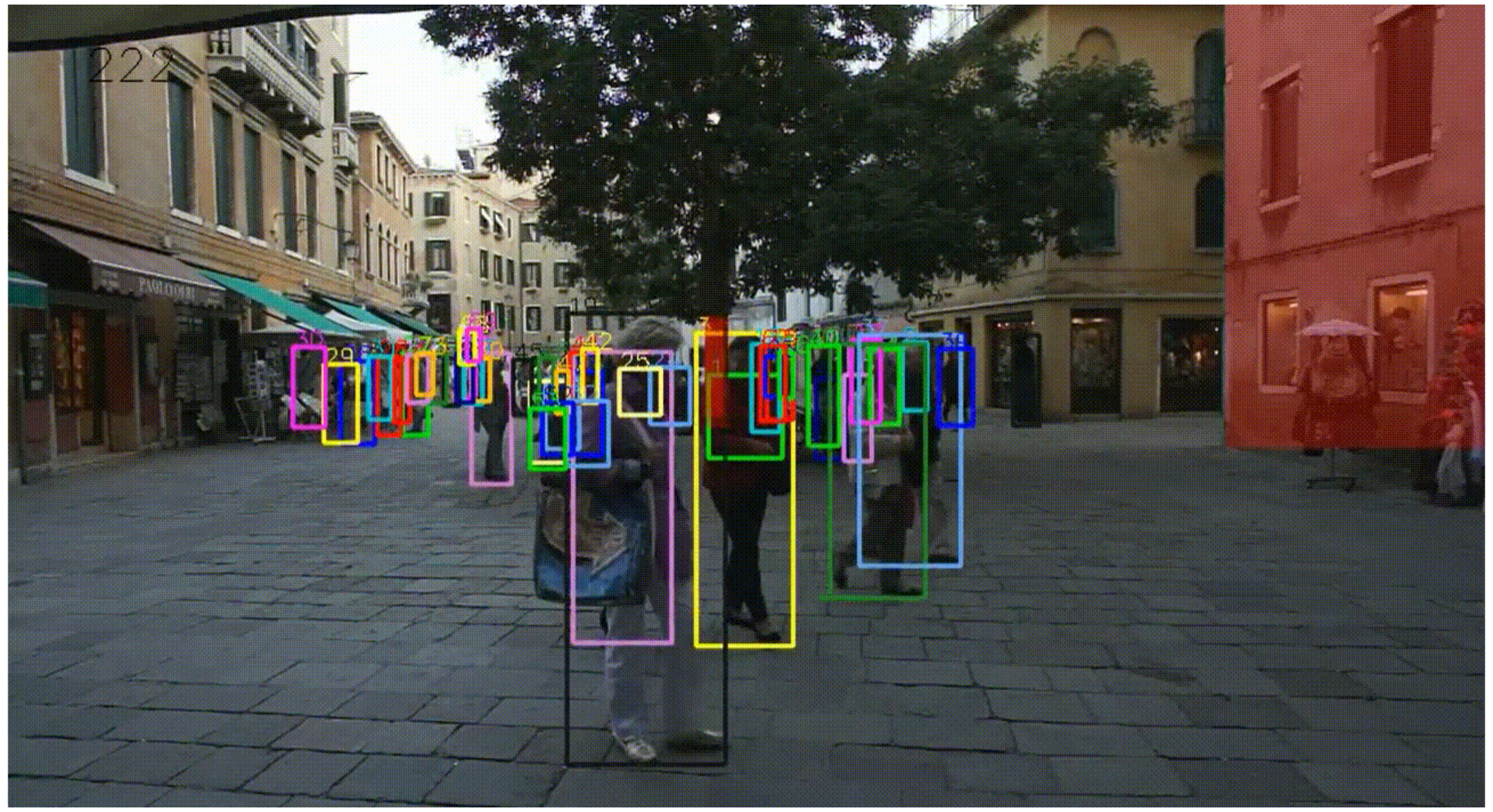
$\mathbf{z}$ continuous and
Linear dynamics

Linear Dynamical System
(LDS)

Dynamical Variational
Auto-encoders
(DVAEs) [1]

[1] Laurent Girin et al., 2021, "Dynamical Variational Autoencoders: A Comprehensive Review", Foundations and Trends in Machine Learning.

Application scenarios: multi-source trajectory separation

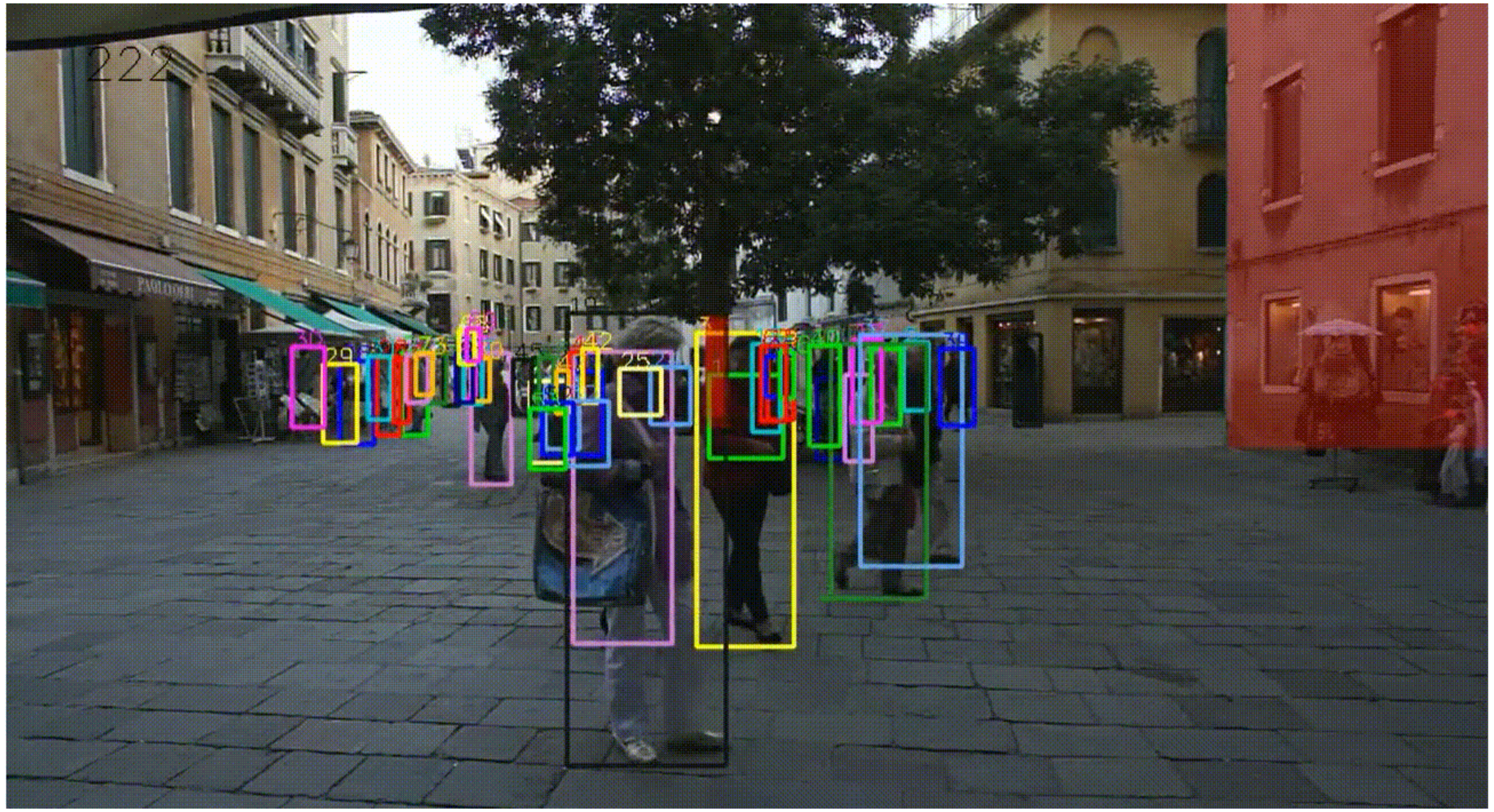


Multi-Object Tracking

Audio Source Separation

Unsupervised multi-object tracking (MOT) with MixDVAE

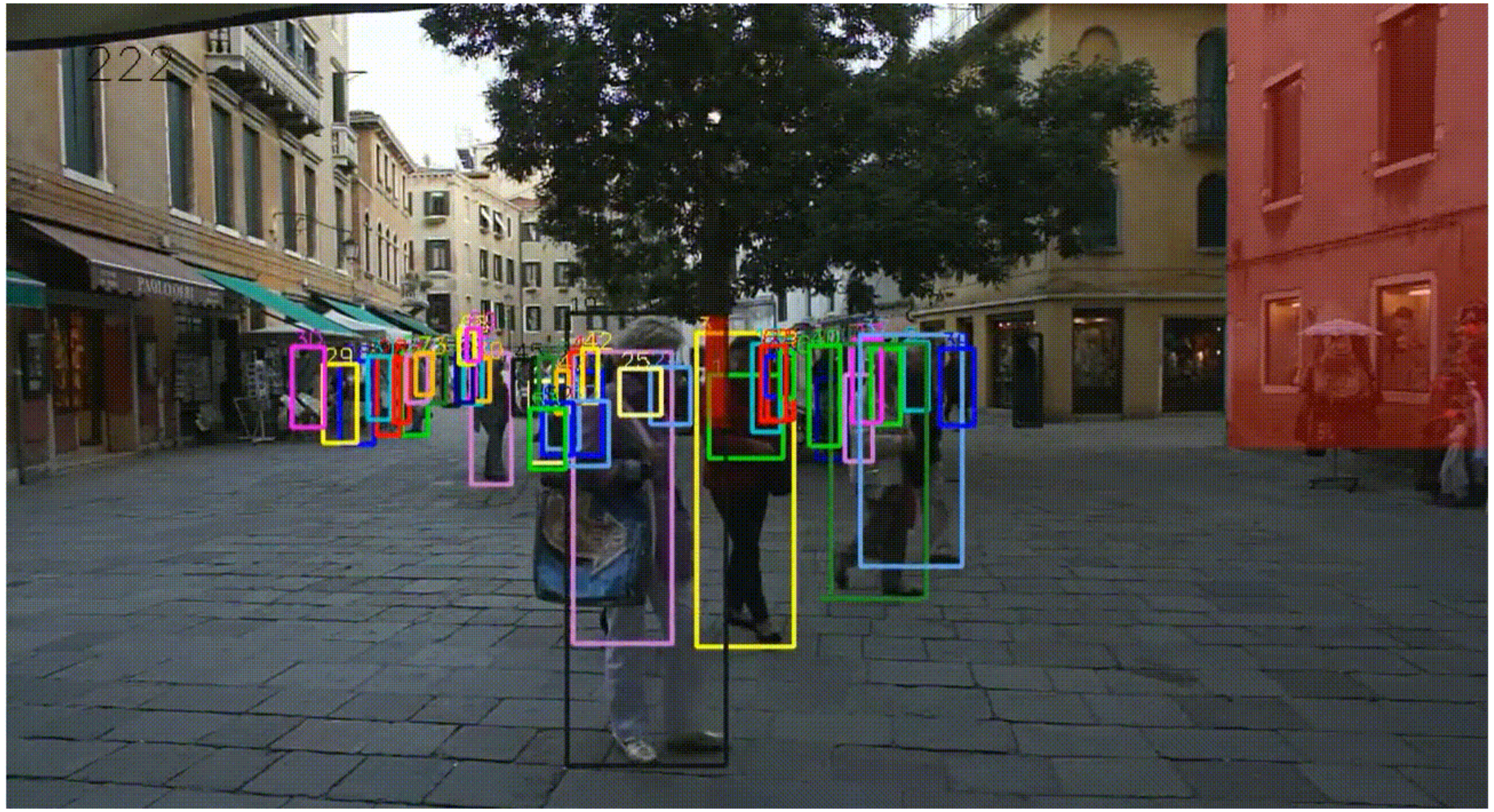
MOT task definition



4 main sub-tasks in MOT

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

Motion-based MOT



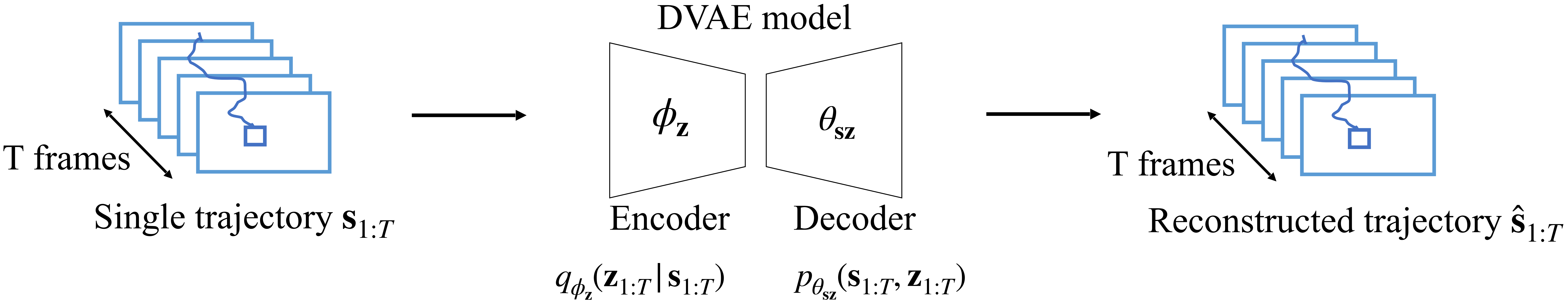
4 main sub-tasks in MOT

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

→ Tracking-by-detection, kown number of sources

Use DVAEs for source motion dynamics modeling

Non-linear probabilistic sequential latent variable generative models



DVAE model

T frames

Single trajectory $\mathbf{s}_{1:T}$

$\phi_{\mathbf{z}}$

$\theta_{\mathbf{sz}}$

Encoder

Decoder

$q_{\phi_{\mathbf{z}}}(\mathbf{z}_{1:T} | \mathbf{s}_{1:T})$

$p_{\theta_{\mathbf{sz}}}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T})$

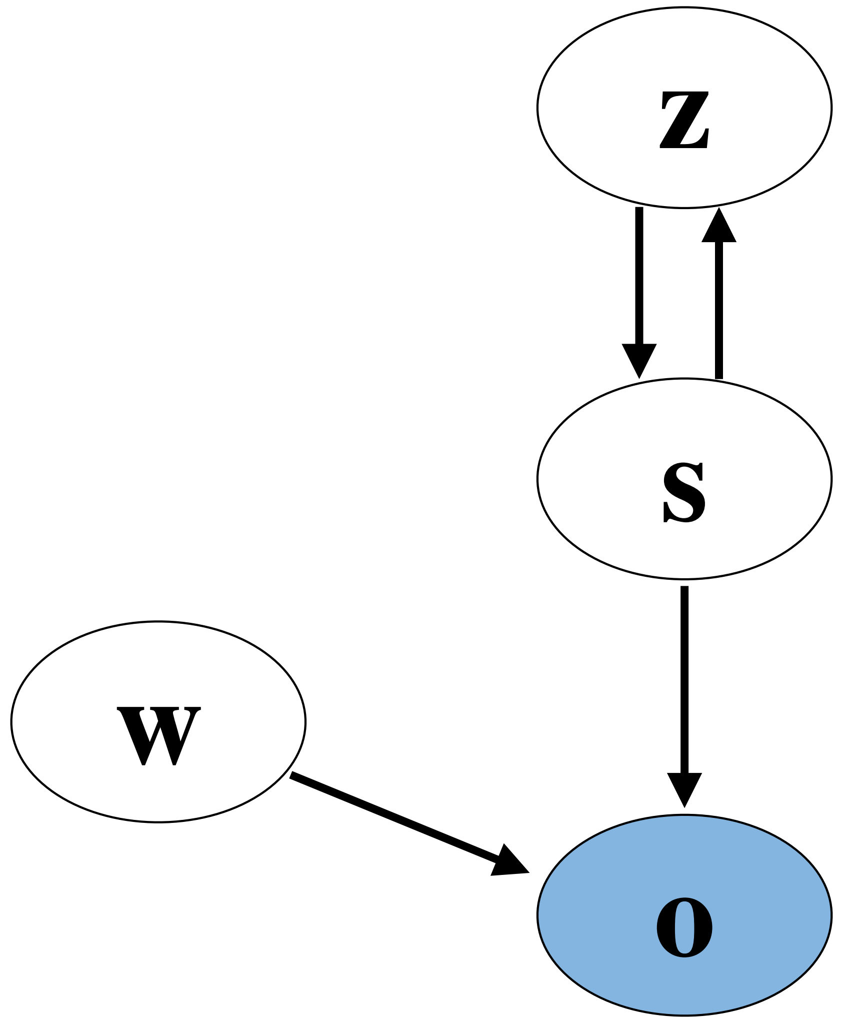
T frames

Reconstructed trajectory $\hat{\mathbf{s}}_{1:T}$

Training by maximizing the Evidence Lower BOund (ELBO)

$$\mathscr{L}(\theta, \phi; \mathbf{s}_{1:T}) = \mathbb{E}_{q_{\phi_{\mathbf{z}}}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})}[\log p_{\theta_{\mathbf{sz}}}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}) - \log q_{\phi_{\mathbf{z}}}(\mathbf{z}_{1:T} | \mathbf{s}_{1:T})]$$

Define MOT from a probabilistic perspective



Definition of random variables

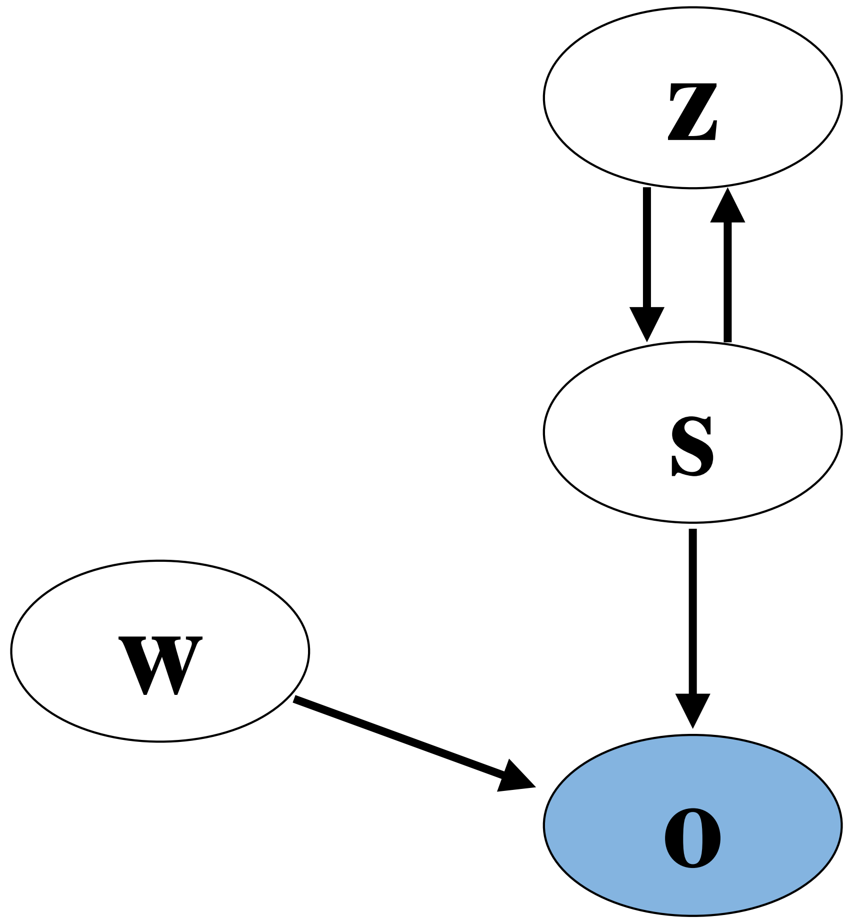
- $\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T\times K_t \times 4}$: positions of detection bounding boxes
- $\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T\times N \times 4}$: true positions of sources
- $\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T\times N \times L}$: latent sequences of DVAE models
- $\mathbf{w} = \{w_{1:T,1:K_t}\} \in \{1,\ldots,N\}^{T\times K_t}$: discrete assignment variables, $w_{tk} = n$ means the observation $\mathbf{o}_{tk}$ is assigned to source n

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

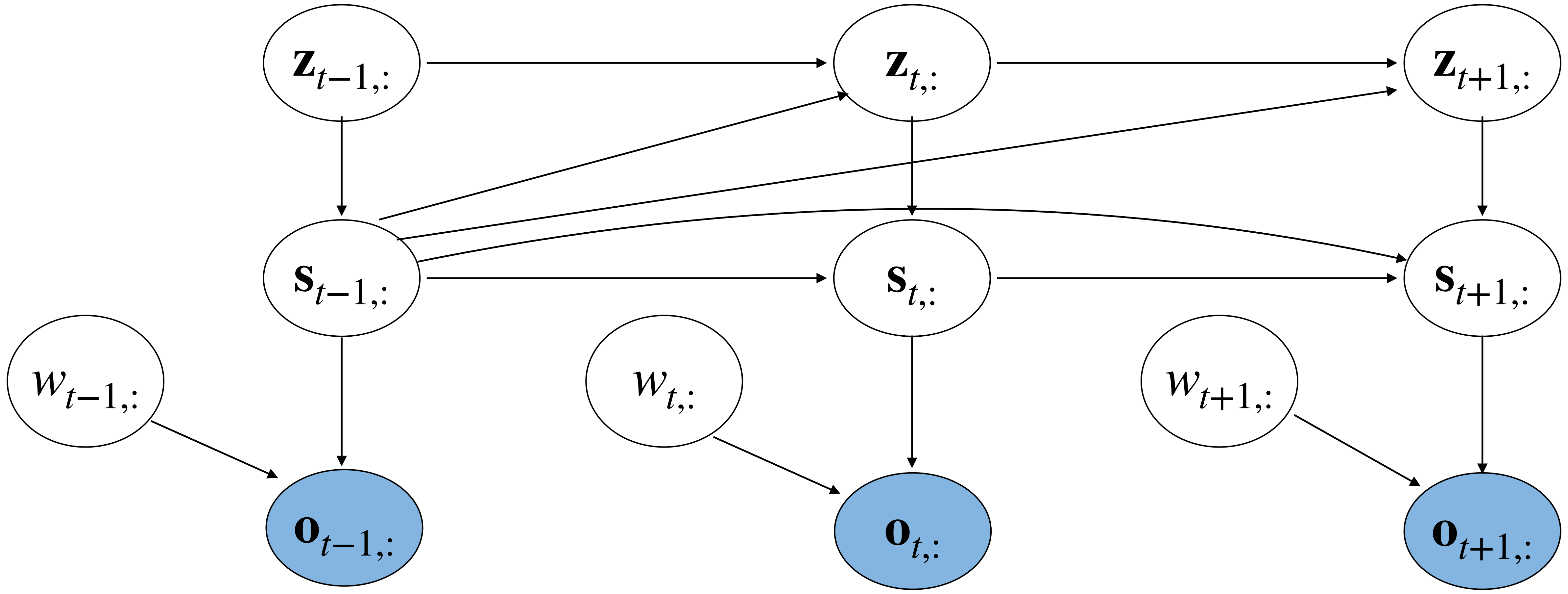
MOT objective: estimate the posterior distribution $p(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

Resolve MOT through Variational Inference (VI)

Associated graphical model



Folded graphical model



Extended graphical model over time frames

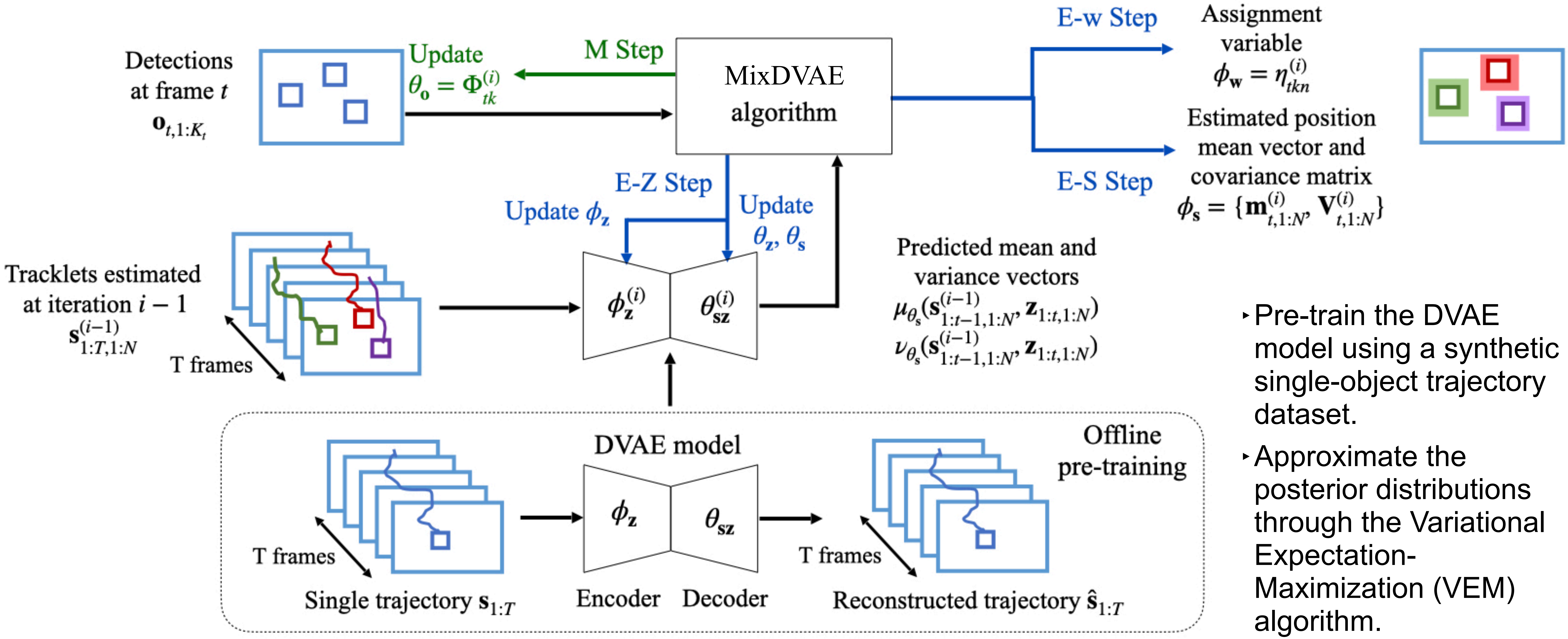
Generative model: $p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o} \,|\, \mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_\mathbf{sz}}(\mathbf{s}, \mathbf{z})$

Intractable true posterior distribution $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \,|\, \mathbf{o})$

Inference model: mean-field like approximation $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \,|\, \mathbf{o}) \approx q_{\phi_\mathbf{w}}(\mathbf{w} \,|\, \mathbf{o}) q_{\phi_\mathbf{z}}(\mathbf{z} \,|\, \mathbf{s}) q_{\phi_\mathbf{s}}(\mathbf{s} \,|\, \mathbf{o})$

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

Resolve MOT through Variational Inference (VI)


Detections at frame t
$\mathbf{o}_{t,1:K_t}$
Update $\theta_\mathbf{o} = \Phi_{tk}^{(i)}$
M Step
MixDVAE algorithm
E-w Step
Assignment variable $\phi_\mathbf{w} = \eta_{tkn}^{(i)}$
E-S Step
Estimated position mean vector and covariance matrix $\phi_\mathbf{s} = \{\mathbf{m}_{t,1:N}^{(i)}, \mathbf{V}_{t,1:N}^{(i)}\}$
E-Z Step
Update $\phi_\mathbf{z}$
Update $\theta_\mathbf{z}, \theta_\mathbf{s}$
Tracklets estimated at iteration $i-1$
$\mathbf{s}_{1:T,1:N}^{(i-1)}$
T frames
$\phi_\mathbf{z}^{(i)}$
$\theta_\mathbf{sz}^{(i)}$
Predicted mean and variance vectors
$\mu_{\theta_\mathbf{s}}(\mathbf{s}_{1:t-1,1:N}^{(i-1)}, \mathbf{z}_{1:t,1:N})$
$\nu_{\theta_\mathbf{s}}(\mathbf{s}_{1:t-1,1:N}^{(i-1)}, \mathbf{z}_{1:t,1:N})$
DVAE model
Offline pre-training
T frames
$\phi_\mathbf{z}$
$\theta_\mathbf{sz}$
T frames
Single trajectory $\mathbf{s}_{1:T}$
Encoder
Decoder
Reconstructed trajectory $\hat{\mathbf{s}}_{1:T}$

- Pre-train the DVAE model using a synthetic single-object trajectory dataset.
- Approximate the posterior distributions through the Variational Expectation-Maximization (VEM) algorithm.

Experimental settings

Datasets

•DVAE pre-training

A synthetic single-source motion trajectories dataset

•Evaluation

MOT17-3T dataset created from the MOT17 training set:

- Subsequences of length T ($T = 60, 120, 300$ frames are tested)

- No birth / death process

- 3 tracking sources per test data sample

Baselines

ArTIST (Saleh et al., 2021), VKF (Ban et al., 2020), Deep AR

Comparison with the SoTA models

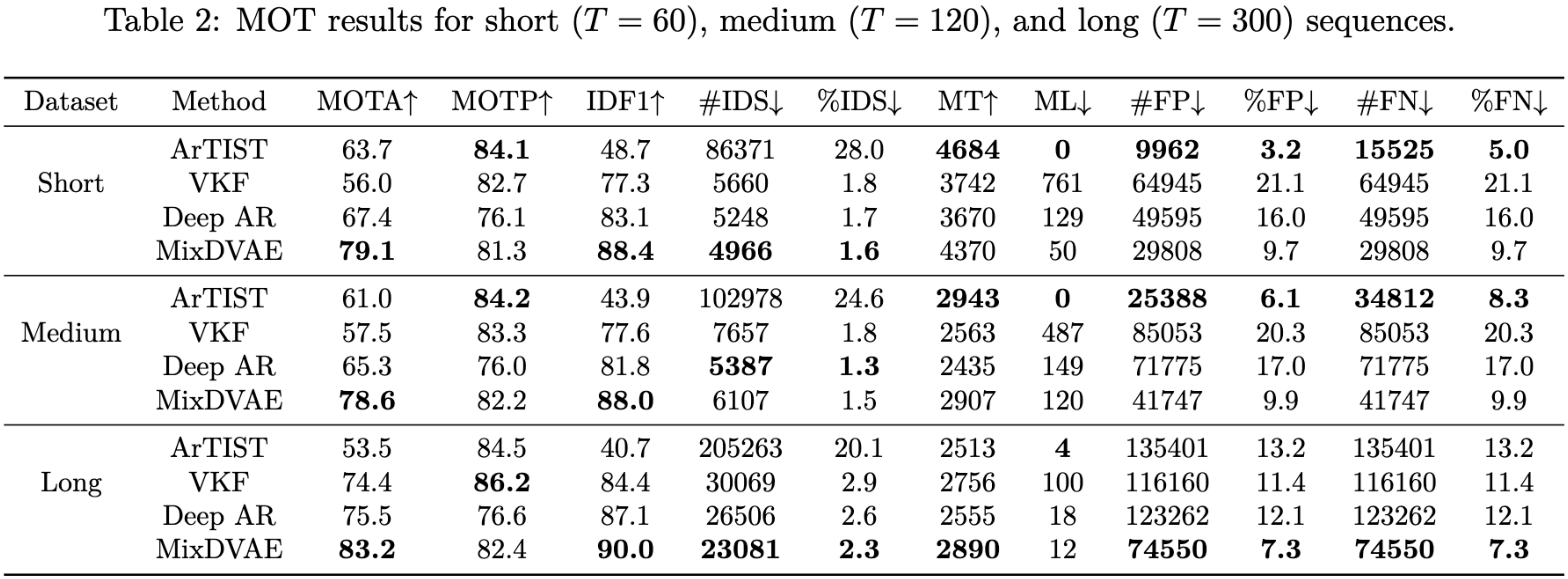
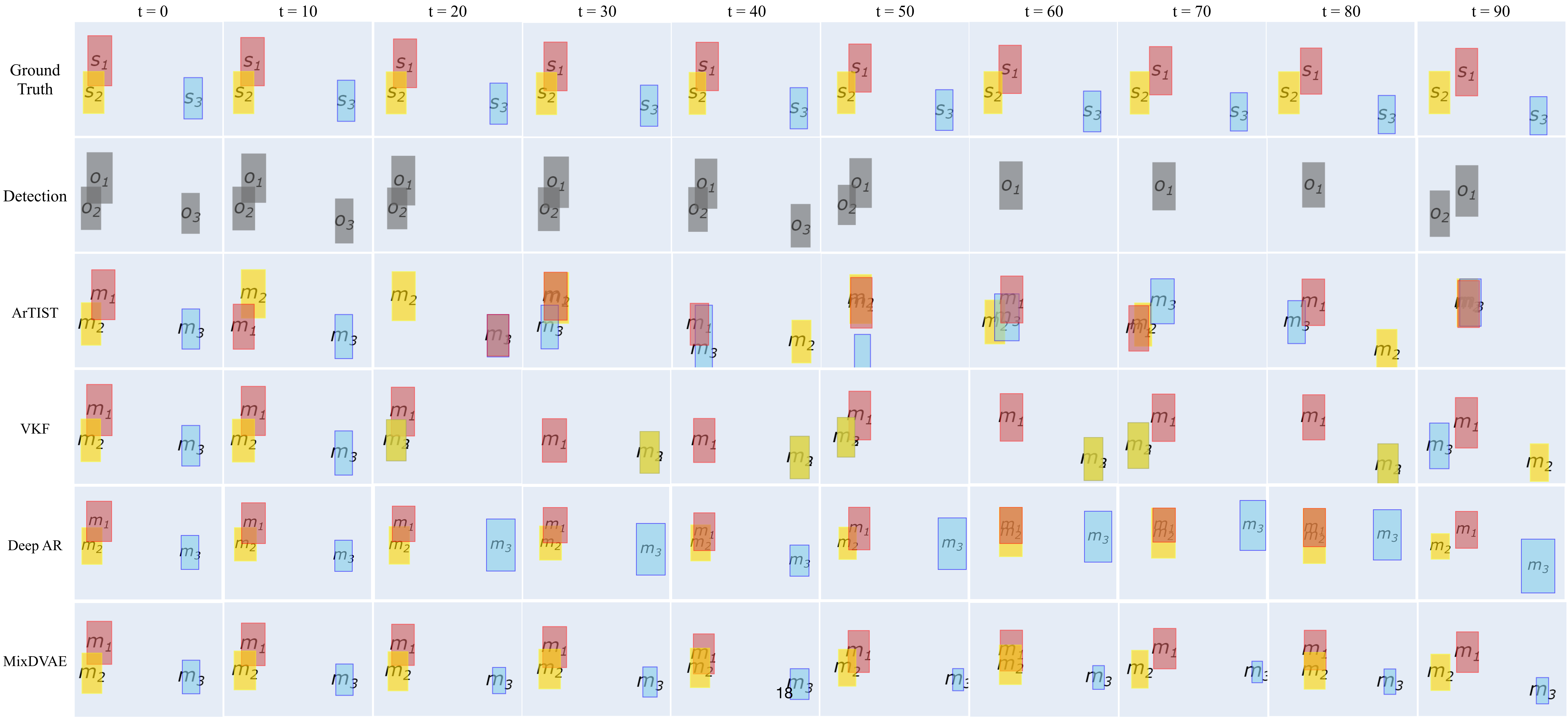


Table 2: MOT results for short ($T = 60$), medium ($T = 120$), and long ($T = 300$) sequences.

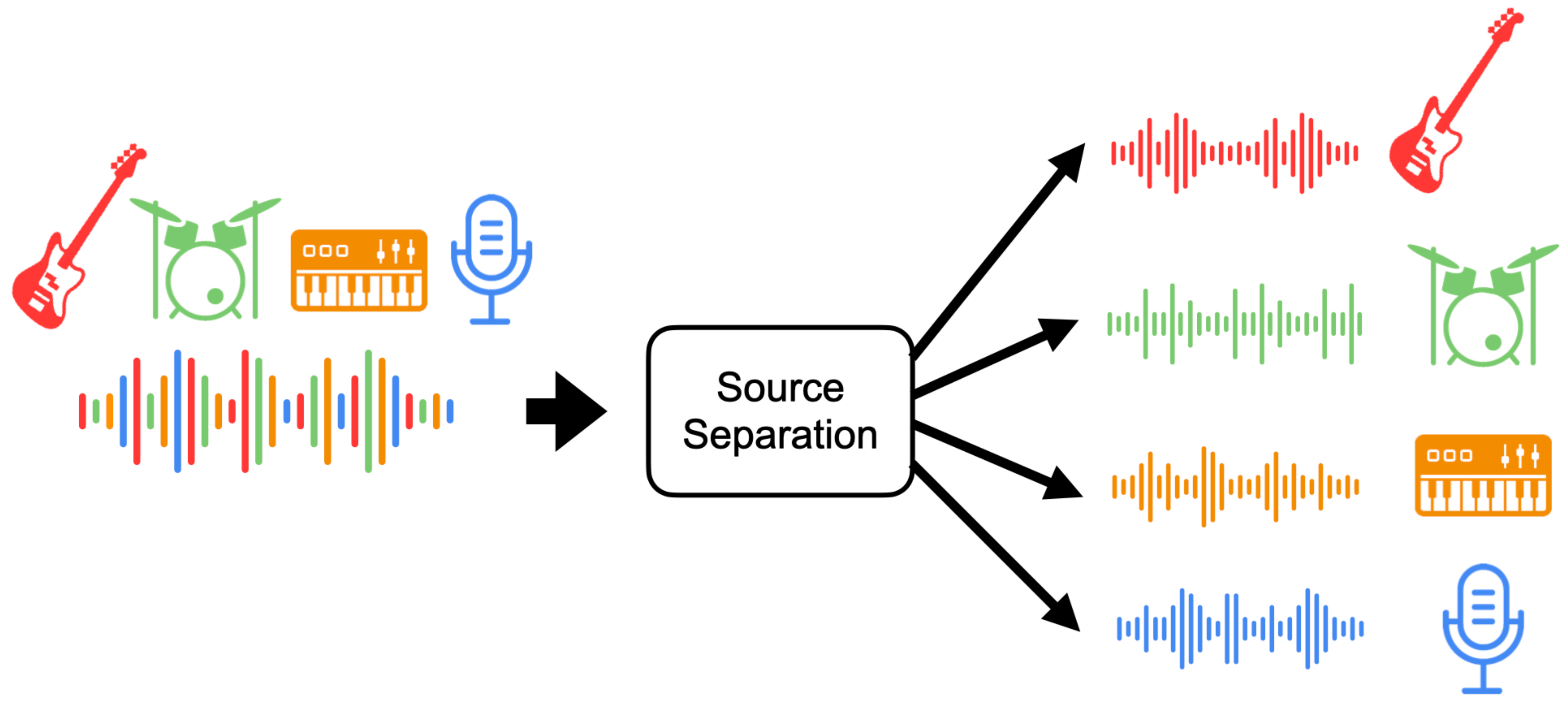
Dataset	Method	MOTA↑	MOTP↑	IDF1↑	#IDS↓	%IDS↓	MT↑	ML↓	#FP↓	%FP↓	#FN↓	%FN↓
Short	ArTIST	63.7	**84.1**	48.7	86371	28.0	**4684**	**0**	**9962**	**3.2**	**15525**	**5.0**
	VKF	56.0	82.7	77.3	5660	1.8	3742	761	64945	21.1	64945	21.1
	Deep AR	67.4	76.1	83.1	5248	1.7	3670	129	49595	16.0	49595	16.0
	MixDVAE	**79.1**	81.3	**88.4**	**4966**	**1.6**	4370	50	29808	9.7	29808	9.7
Medium	ArTIST	61.0	**84.2**	43.9	102978	24.6	**2943**	**0**	**25388**	**6.1**	**34812**	**8.3**
	VKF	57.5	83.3	77.6	7657	1.8	2563	487	85053	20.3	85053	20.3
	Deep AR	65.3	76.0	81.8	**5387**	**1.3**	2435	149	71775	17.0	71775	17.0
	MixDVAE	**78.6**	82.2	**88.0**	6107	1.5	2907	120	41747	9.9	41747	9.9
Long	ArTIST	53.5	84.5	40.7	205263	20.1	2513	**4**	135401	13.2	135401	13.2
	VKF	74.4	**86.2**	84.4	30069	2.9	2756	100	116160	11.4	116160	11.4
	Deep AR	75.5	76.6	87.1	26506	2.6	2555	18	123262	12.1	123262	12.1
	MixDVAE	**83.2**	82.4	**90.0**	**23081**	**2.3**	**2890**	12	**74550**	**7.3**	**74550**	**7.3**

Tracking example visualization



Weakly supervised single-channel audio source separation with MixDVAE

Audio source separation

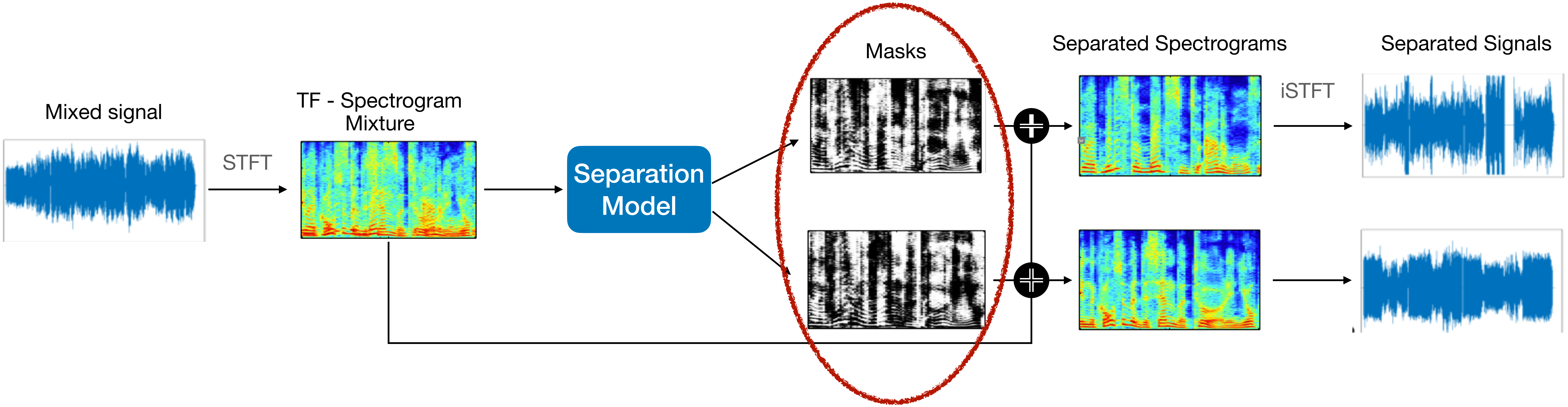


“Cocktail Party Effect” — Bregman 1990

Applications

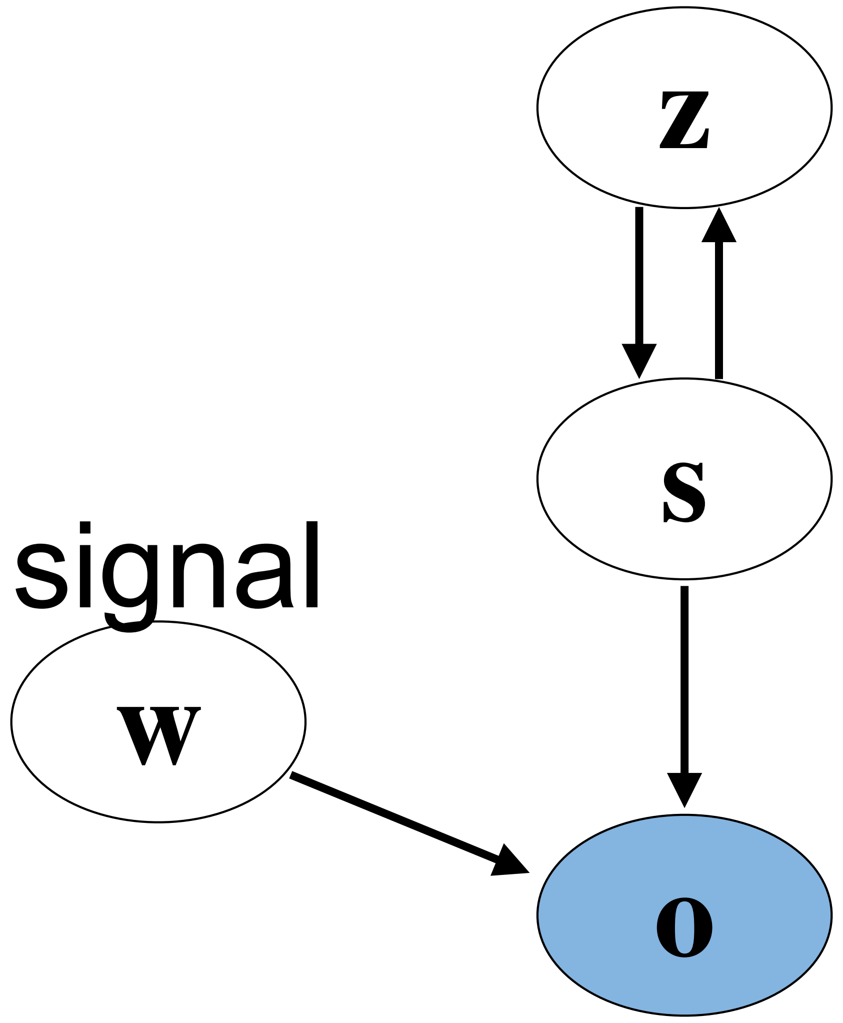
- real-time speaker separation
- speech enhancement within hearing aids
- voice cancellation for karaoke
- …

SC-ASS: Time-Frequency Masking with probabilistic models



Key question: how to obtain the masks?

Define SC-ASS from a probabilistic perspective



Definition of random variables

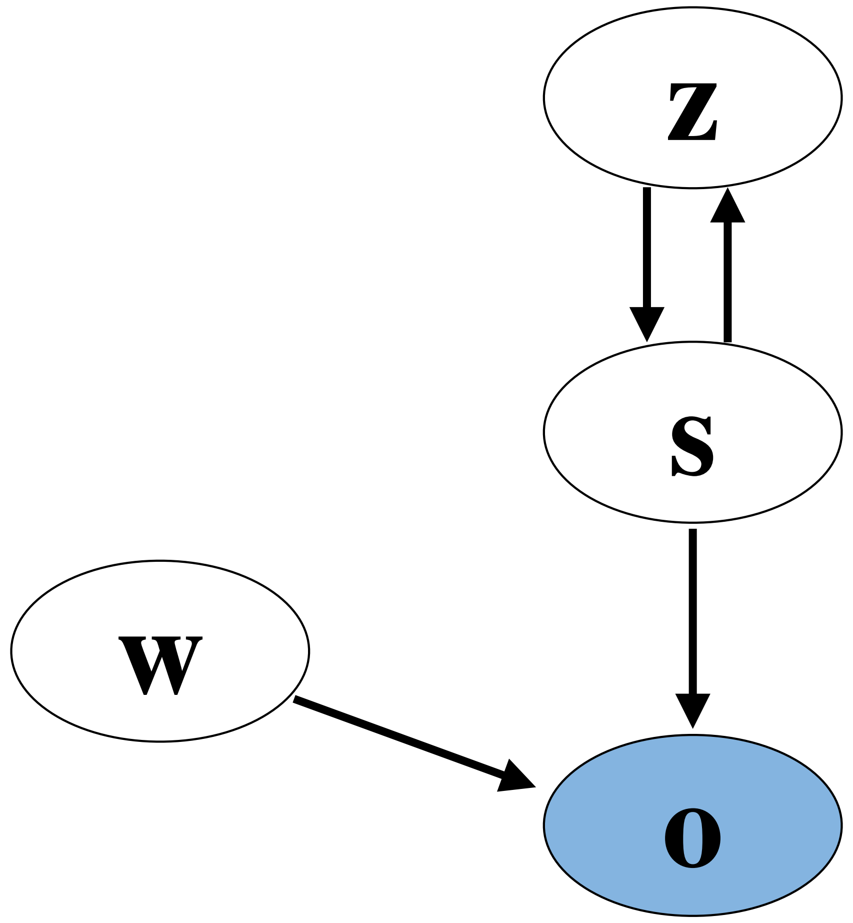
- $\mathbf{o} = \{o_{1:T,1:F}\} \in \mathbb{C}^{T \times F}$: STFT spectrogram of the observed mixture signal
- $\mathbf{s} = \{s_{1:N,1:T,1:F}\} \in \mathbb{C}^{N \times T \times F}$: STFT spectrograms of N sources
- $\mathbf{z} = \{\mathbf{z}_{1:N,1:T}\} \in \mathbb{R}^{N \times T \times L}$: latent sequences of DVAE models
- $\mathbf{w} = \{w_{1:T,1:F}\} \in \{1,\ldots,N\}^{T \times F}$: discrete assignment variables, $w_{tf} = n$ means the mixture signal at TF bin [t, f] $o_{t,f}$ is assigned to source n

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

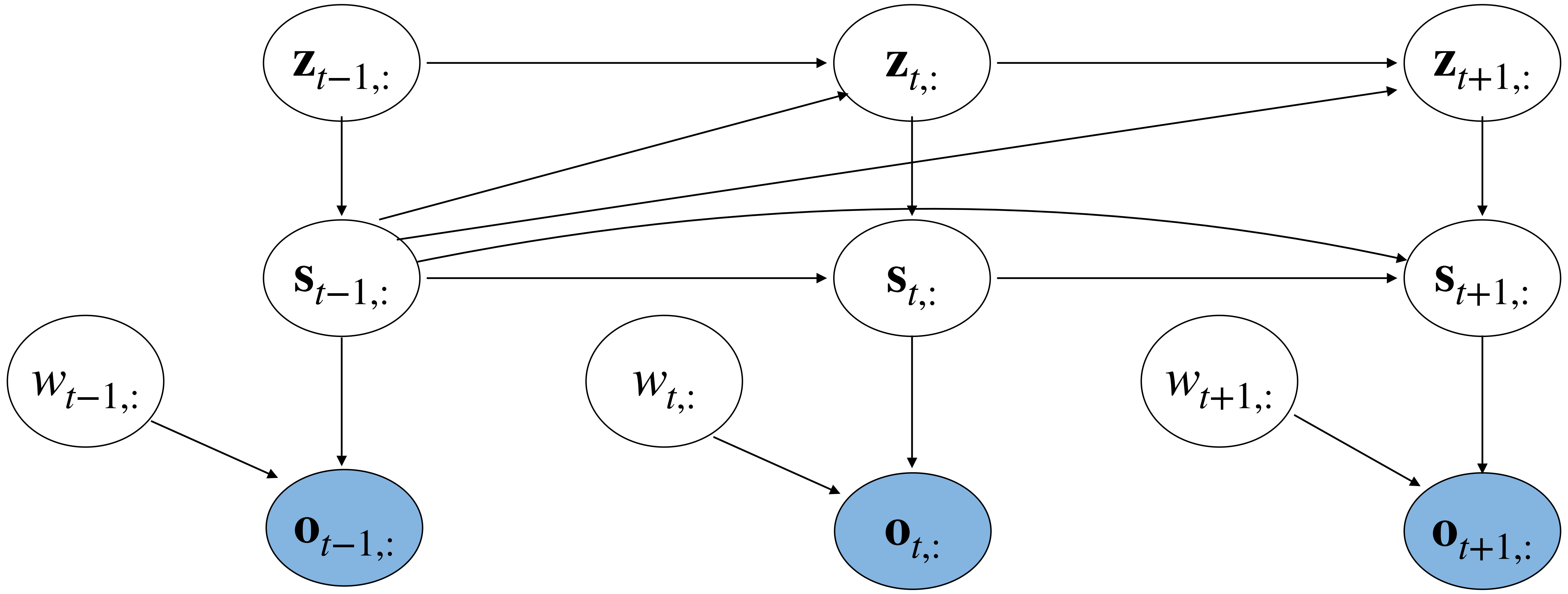
SC-ASS objective: estimate the posterior distribution $p(\mathbf{s}, \mathbf{z}, \mathbf{w} \,|\, \mathbf{o})$

Resolve SC-ASS through Variational Inference (VI)

Associated graphical model



Folded graphical model



Extended graphical model over time frames

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

These distributions are different from that of the MOT problem.

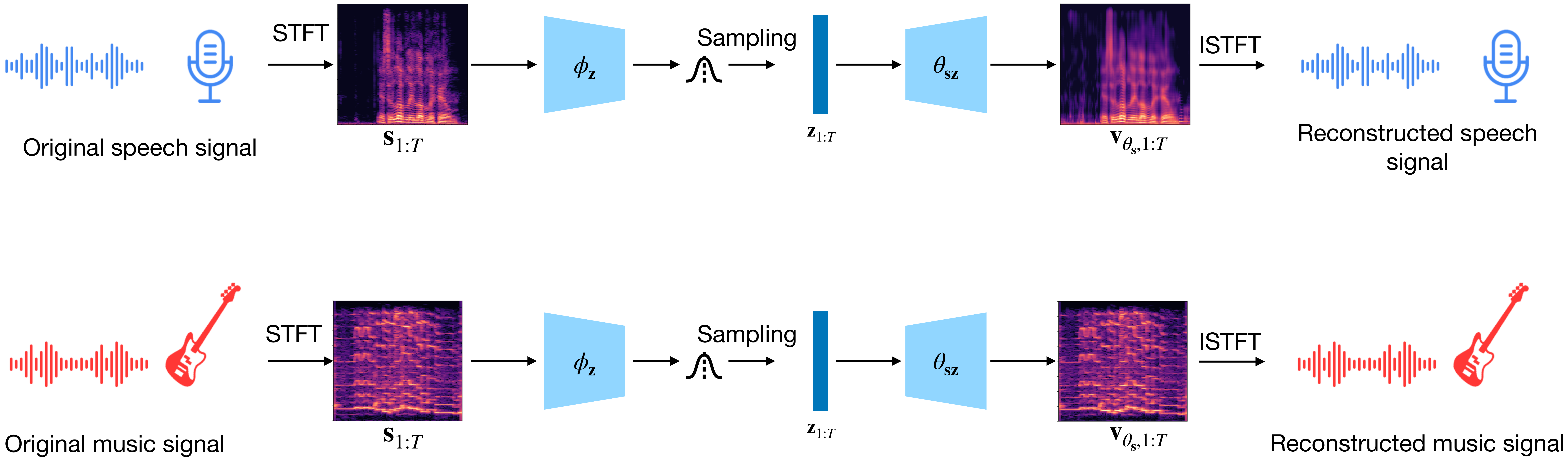
Intractable true posterior distribution $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

Inference model: mean-field like approximation $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o}) \approx q_{\phi_\mathbf{w}}(\mathbf{w} \mid \mathbf{o}) q_{\phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{s}) q_{\phi_\mathbf{s}}(\mathbf{s} \mid \mathbf{o})$

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

Resolve SC-ASS through Variational Inference (VI)

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



Resolve SC-ASS through Variational Inference (VI)

Mixture spectrogram $\mathbf{o} \in \mathbb{C}^{T\times F}$
Update $\Phi^{(i)}_{tk}$
M Step
MixDVAE algorithm
E-W Step
Assignment variable w_{tf}
Estimated source spectrograms $\mathbf{s}_{1:N} \in \mathbb{C}^{N\times T\times F}$
E-S Step
E-Z Step
Update $\phi_{\mathbf{z}}$
Update $\theta_{\mathbf{z}}, \theta_{\mathbf{s}}$
Initialization: mixture spectrogram $\mathbf{o} \in \mathbb{C}^{T\times F}$
$\phi^{(i)}_{\mathbf{z}}$
$\theta^{(i)}_{\mathbf{sz}}$
DVAE predicted spectrograms
$\mathbf{v}_{\theta_{\mathbf{s}},\mathbf{1:N}}(\mathbf{s}^{(i-1)}_{1:t-1}, \mathbf{z}_{1:t})$
DVAE model
$\phi_{\mathbf{z}}$
$\theta_{\mathbf{sz}}$
Offline pre-training

Experimental settings

Datasets

- DVAE pre-training
  - Wall Street Journal (WSJ0) dataset (Garofolo et al., 1993)
  - Chinese Bamboo Flute (CBF) dataset (Wang et al., 2022)
- Evaluation

Mixture signal 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, Deep AR, MixIT (Wisdom et al., 2020), Vanilla NMF (Févotte et al., 2018), temporal NMF (Virtanen, 2007)

Comparison with baseline models

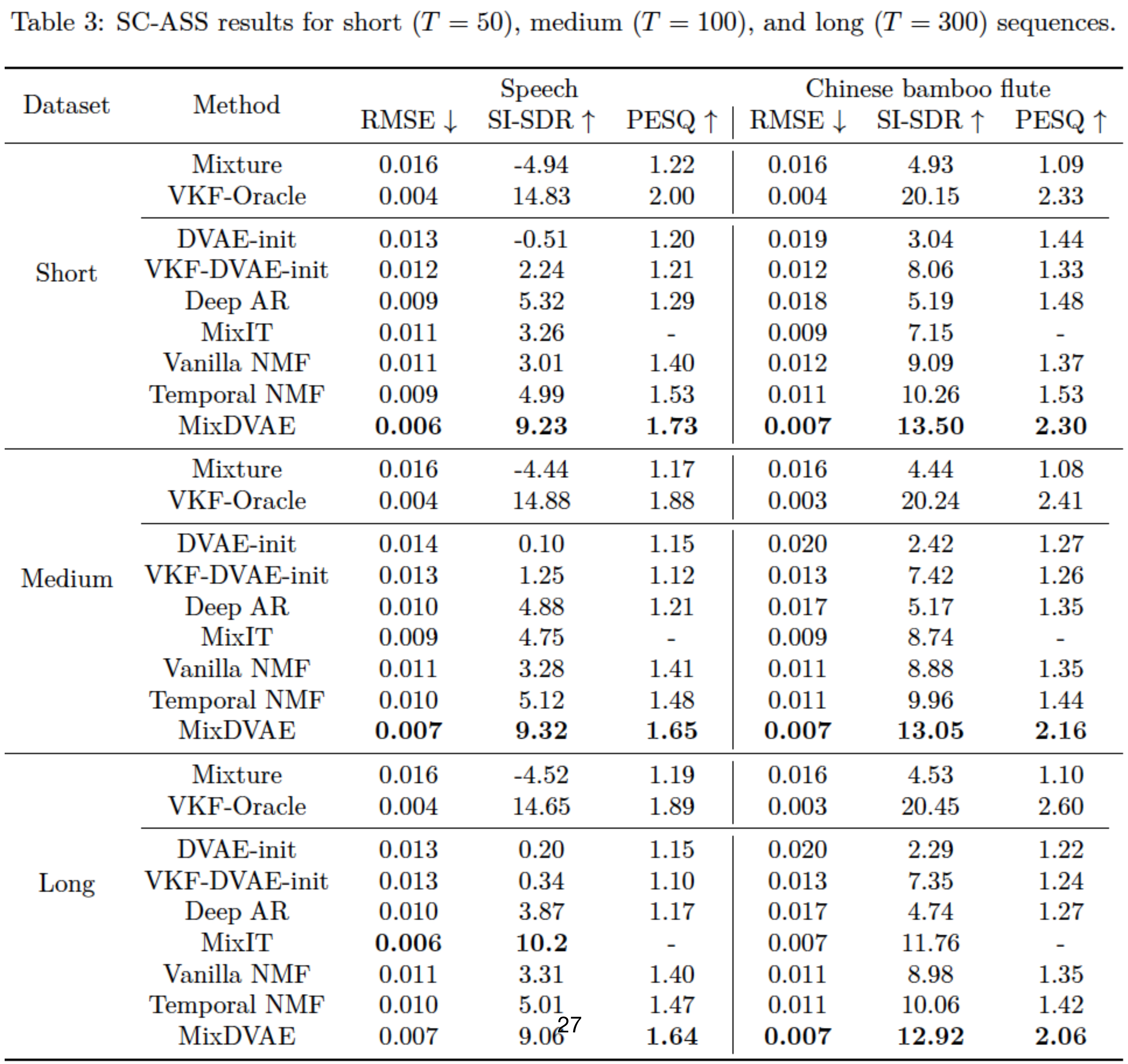
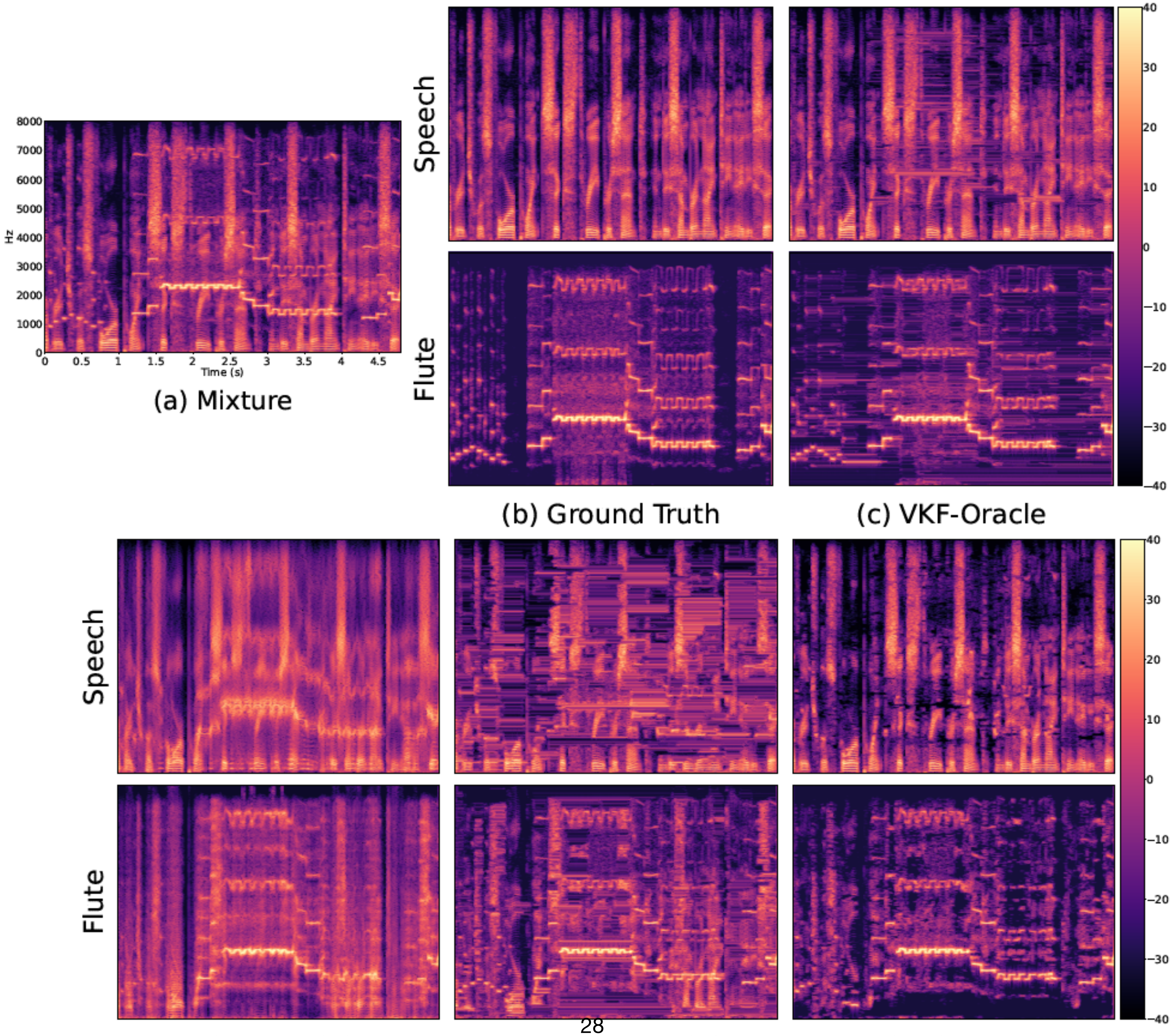


Table 3: SC-ASS results for short ($T = 50$), medium ($T = 100$), and long ($T = 300$) sequences.

Dataset	Method	Speech			Chinese bamboo flute		
		RMSE ↓	SI-SDR ↑	PESQ ↑	RMSE ↓	SI-SDR ↑	PESQ ↑
Short	Mixture	0.016	-4.94	1.22	0.016	4.93	1.09
	VKF-Oracle	0.004	14.83	2.00	0.004	20.15	2.33
	DVAE-init	0.013	-0.51	1.20	0.019	3.04	1.44
	VKF-DVAE-init	0.012	2.24	1.21	0.012	8.06	1.33
	Deep AR	0.009	5.32	1.29	0.018	5.19	1.48
	MixIT	0.011	3.26	-	0.009	7.15	-
	Vanilla NMF	0.011	3.01	1.40	0.012	9.09	1.37
	Temporal NMF	0.009	4.99	1.53	0.011	10.26	1.53
	MixDVAE	**0.006**	**9.23**	**1.73**	**0.007**	**13.50**	**2.30**
Medium	Mixture	0.016	-4.44	1.17	0.016	4.44	1.08
	VKF-Oracle	0.004	14.88	1.88	0.003	20.24	2.41
	DVAE-init	0.014	0.10	1.15	0.020	2.42	1.27
	VKF-DVAE-init	0.013	1.25	1.12	0.013	7.42	1.26
	Deep AR	0.010	4.88	1.21	0.017	5.17	1.35
	MixIT	0.009	4.75	-	0.009	8.74	-
	Vanilla NMF	0.011	3.28	1.41	0.011	8.88	1.35
	Temporal NMF	0.010	5.12	1.48	0.011	9.96	1.44
	MixDVAE	**0.007**	**9.32**	**1.65**	**0.007**	**13.05**	**2.16**
Long	Mixture	0.016	-4.52	1.19	0.016	4.53	1.10
	VKF-Oracle	0.004	14.65	1.89	0.003	20.45	2.60
	DVAE-init	0.013	0.20	1.15	0.020	2.29	1.22
	VKF-DVAE-init	0.013	0.34	1.10	0.013	7.35	1.24
	Deep AR	0.010	3.87	1.17	0.017	4.74	1.27
	MixIT	**0.006**	**10.2**	-	0.007	11.76	-
	Vanilla NMF	0.011	3.31	1.40	0.011	8.98	1.35
	Temporal NMF	0.010	5.01	1.47	0.011	10.06	1.42
	MixDVAE	0.007	9.06	**1.64**	**0.007**	**12.92**	**2.06**

SC-ASS example visualization



(a) Mixture

(b) Ground Truth

(c) VKF-Oracle

(d) DVAE-init

(e) VKF-DVAE-init

(f) MixDVAE

Conclusions

Advantages

- Data efficiency: no need for large amount of annotated data
- Interpretability
- Prediction uncertainty calibration

Limitations

- Computational efficiency

Further discussions

Context

- Boom of large models trained over large datasets: generative models, foundation models.
- Practical concerns about model transparency, interpretability, uncertainty calibration, data efficiency, and human-model interaction.

Open question

- How can statistical and probabilistic knowledge be effectively integrated into DL architectures to enhance the design of more robust models?

Evaluation factors

- Performance
- Computation efficiency
- Generalization ability
- …

→

New Learning framework

- Training configurations: un/semi/self-supervision
- Optimization methods
- Model design
- ….

Q & A