Deep Probabilistic Generative Models for Audio/Visual tasks

Xiaoyu LIN
INRIA, Univ. Grenoble-Alpes

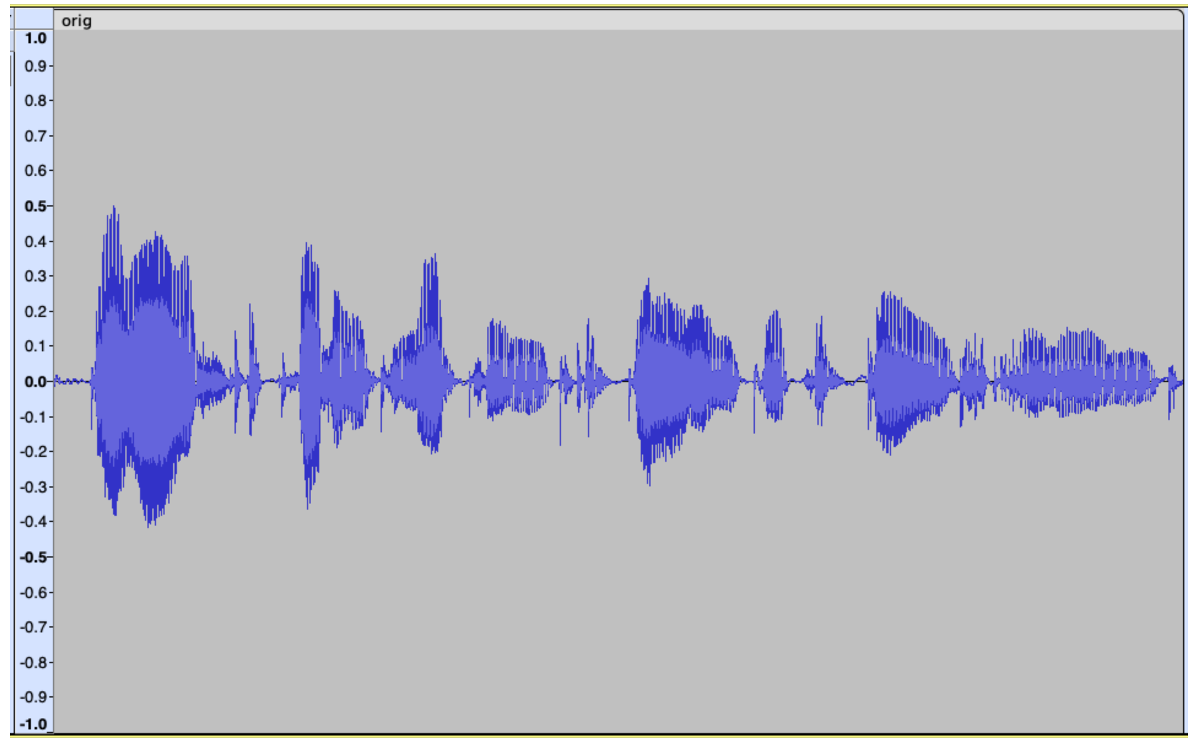
November 21, 2023

Probabilistic Generative Models

Motivations

- Understand complex real-world data

Image

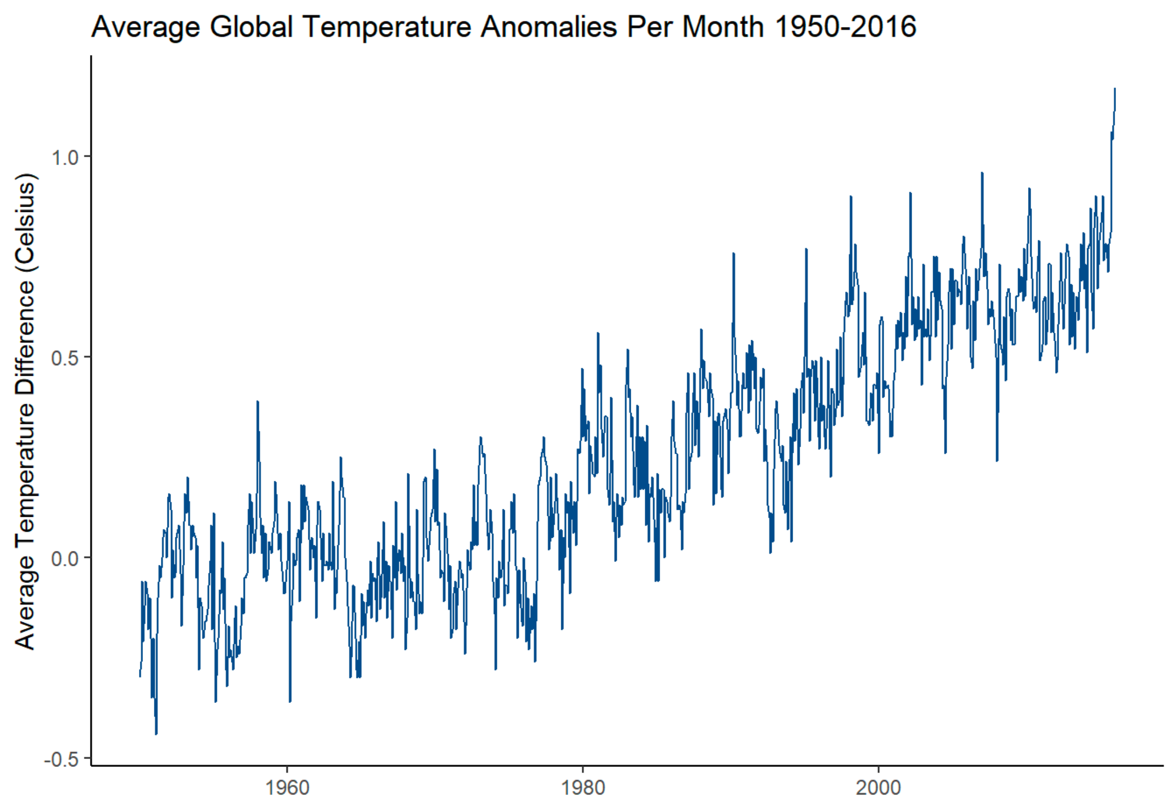


Audio

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

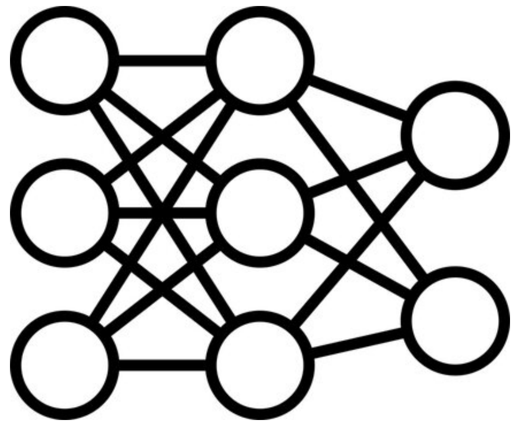


Time series

Motivations

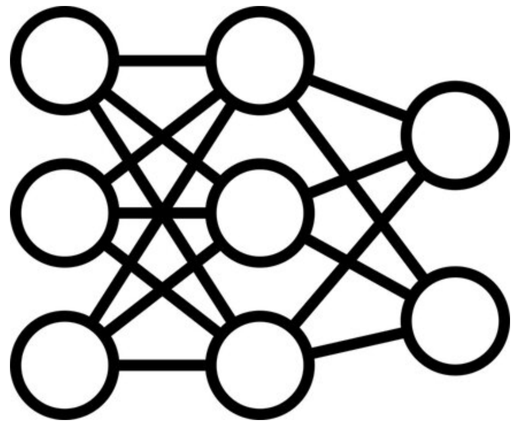
- Understand complex real-world data
- Generate new data points

"An astronaut riding a horse"

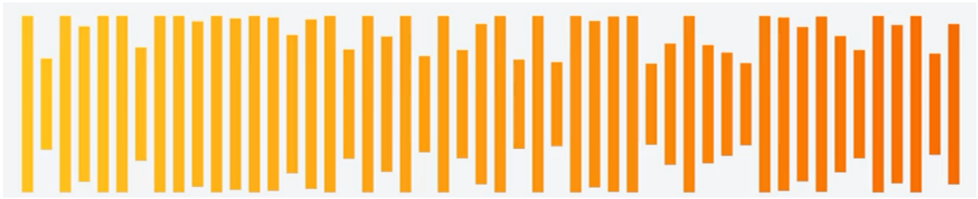


generative model

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



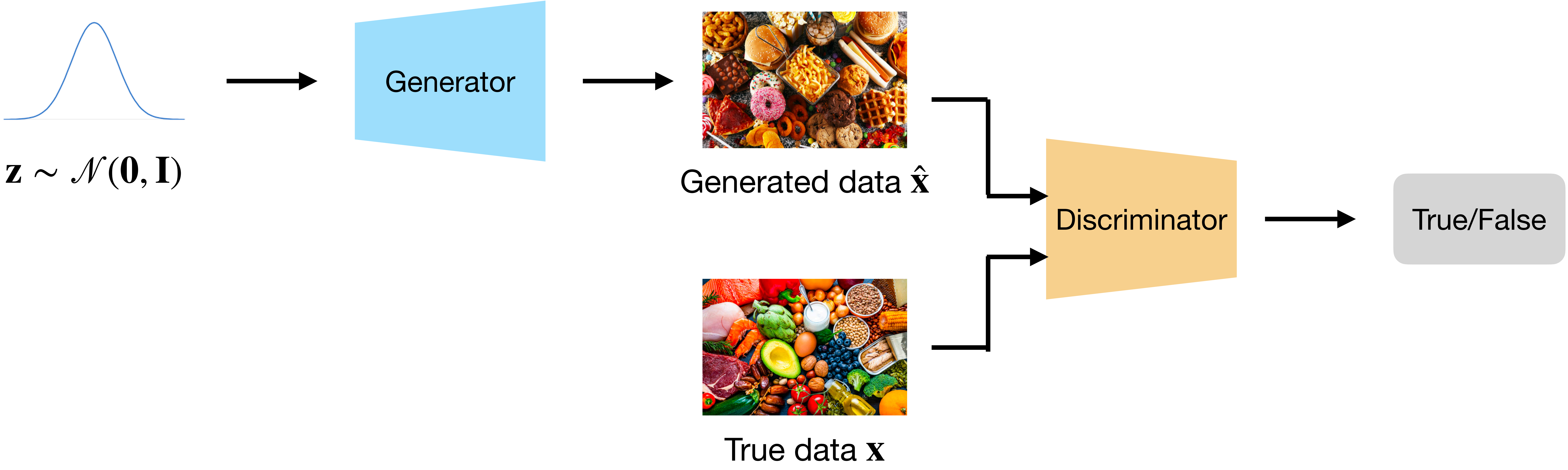
generative model



Examples from DALLE 2 and MusicGen

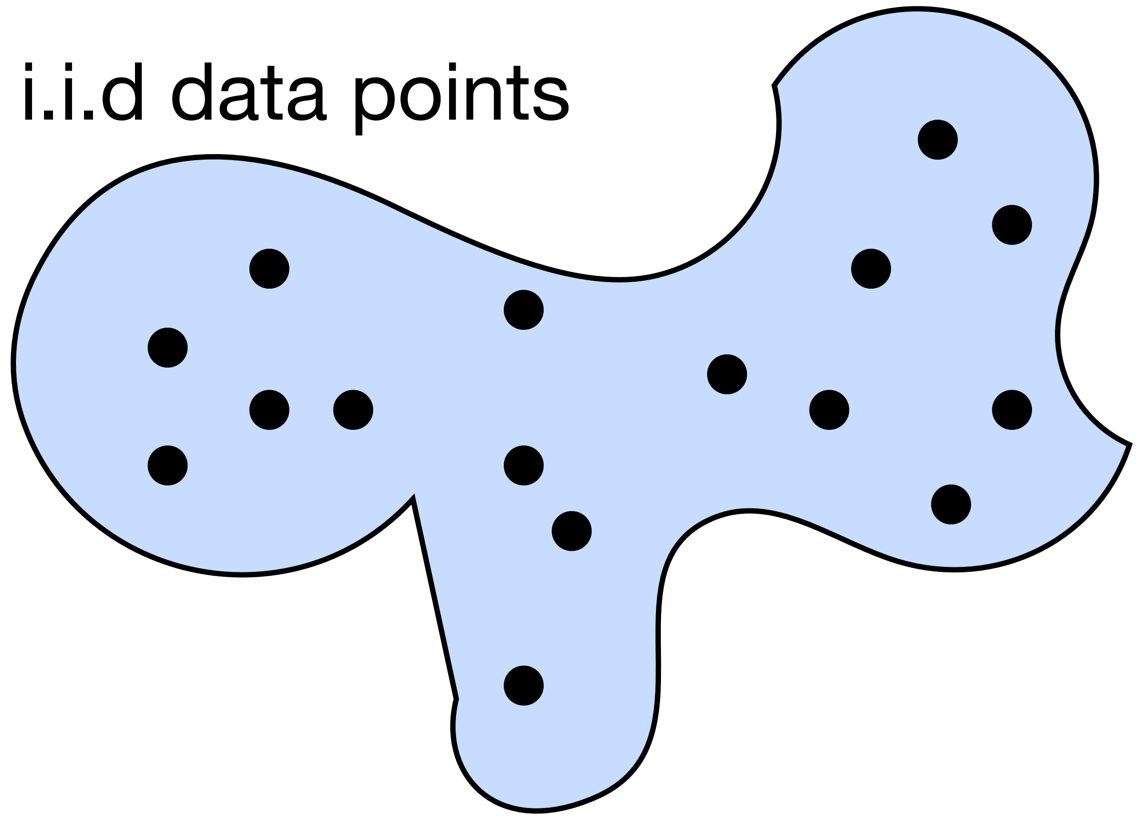
Approaches

- Implicit generative models
  - Generative Adversarial Networks (GANs)



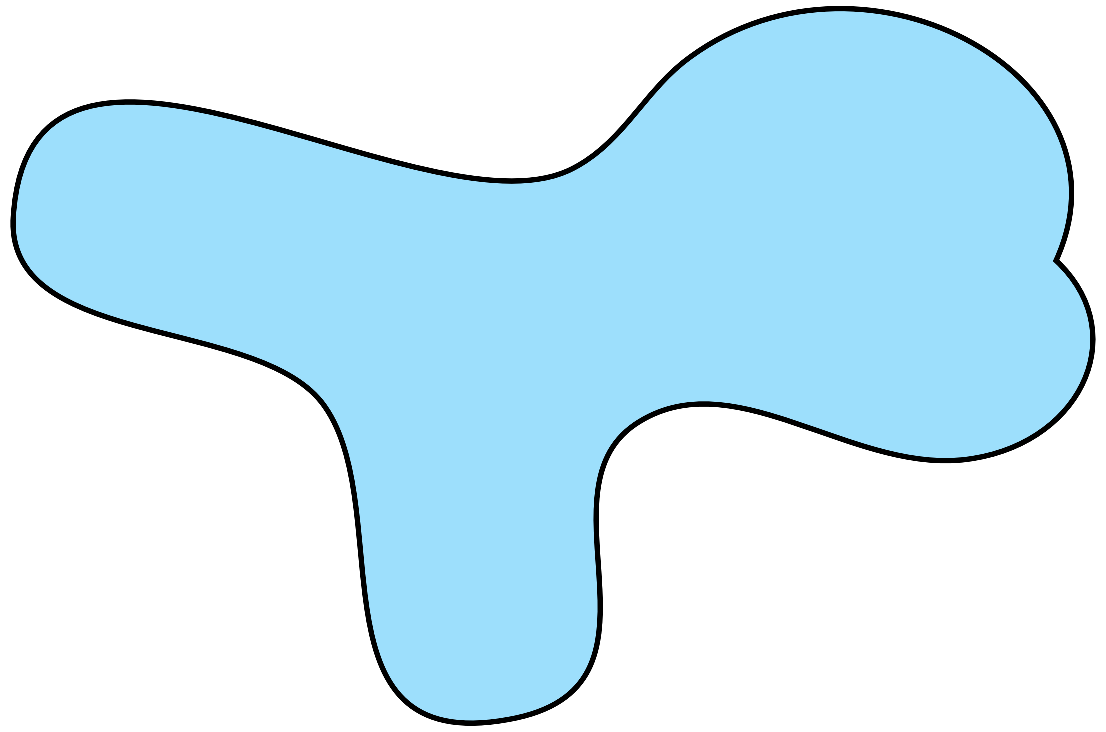
Approaches

- 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

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

Approaches

- Implicit generative models
  - Generative Adversarial Networks (GANs)
- Explicit generative models
  - Auto-regressive models: $p_\theta(\mathbf{x}) = \prod_{i=1}^{d} p_\theta(x_i \mid \mathbf{x}_{<i})$
  - Energy-based models: $p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z_\theta}$
  - Score-based models: $s_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$
  - Normalizing flows: $\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}),\ \mathbf{x} = f_\theta(\mathbf{z}),\ p_\theta(\mathbf{x}) = p_{\mathbf{z}}(f_\theta^{-1}(\mathbf{x})) \mid \det(\mathbf{J}_{f_\theta^{-1}}(\mathbf{x})) \mid$
  - Diffusion models: $\mathbf{x}_0 \sim p_{data}(\mathbf{x}),\ \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ p_\theta(\mathbf{x}_0) = \int p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \mathrm{d}\mathbf{x}_{1:T}$
  - $\boxed{\text{Latent variable models: } p_\theta(\mathbf{x}) = \int p(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z}) \mathrm{d}\mathbf{z}}$

Propose a specific form of $p_\theta(\mathbf{x})$

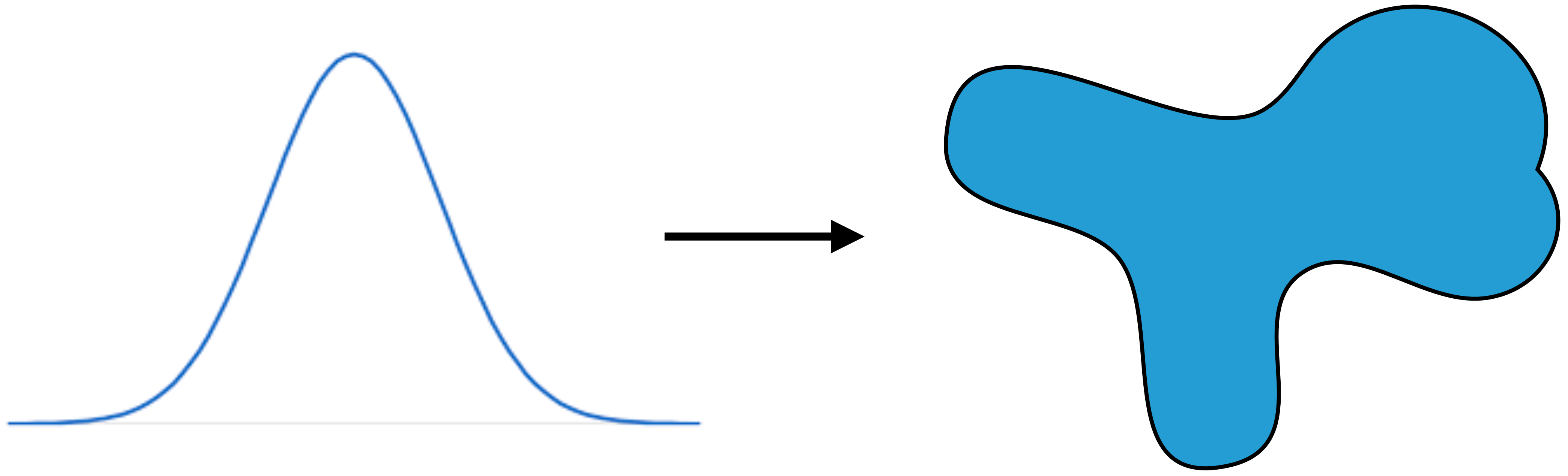
Construct $p_\theta(\mathbf{x})$ from a known simple distribution

Latent Variable Models and Variational Inference

Two main objectives of latent variable models

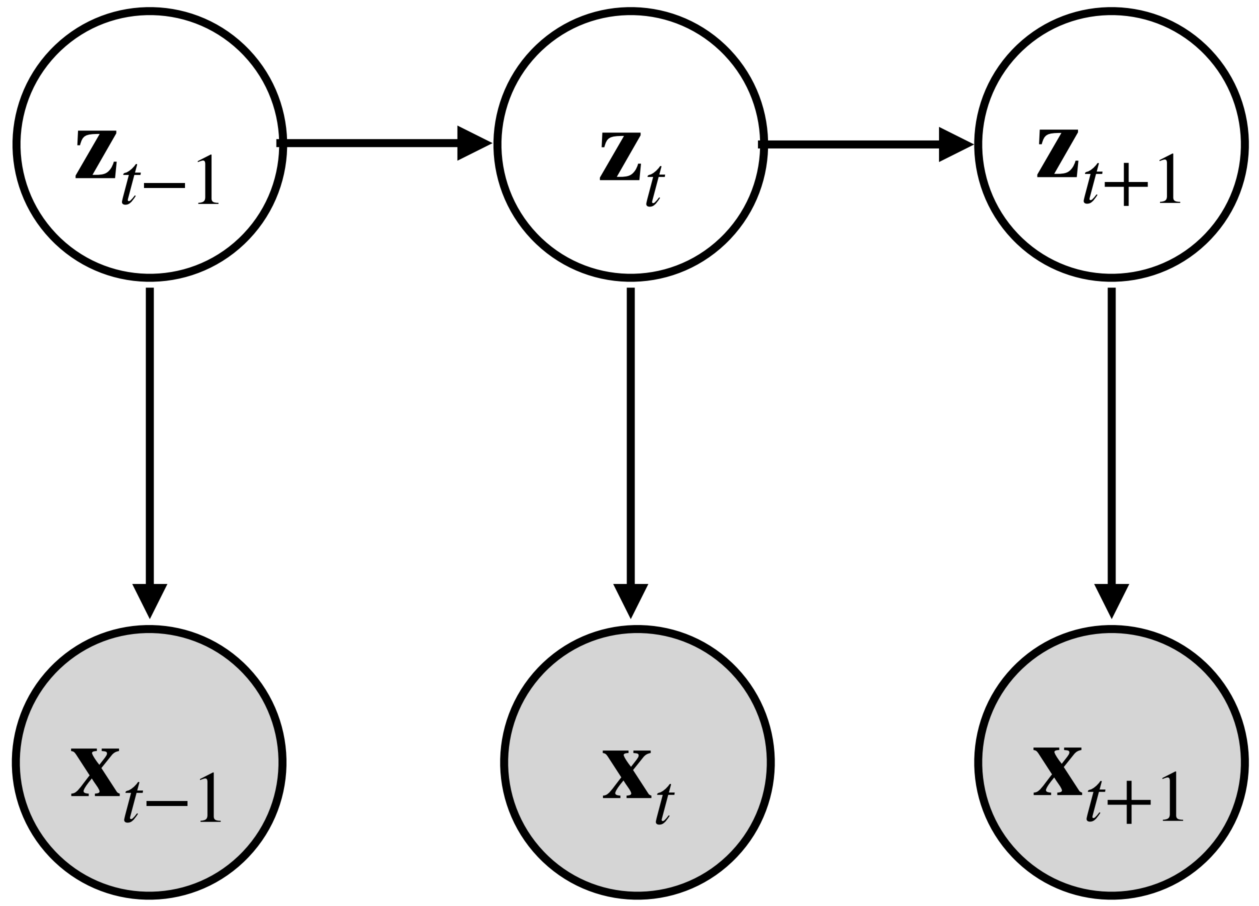
- Help to construct more complex distributions

$$p_\theta(x) = \int p_\theta(x \mid z) p_\theta(z) dz$$

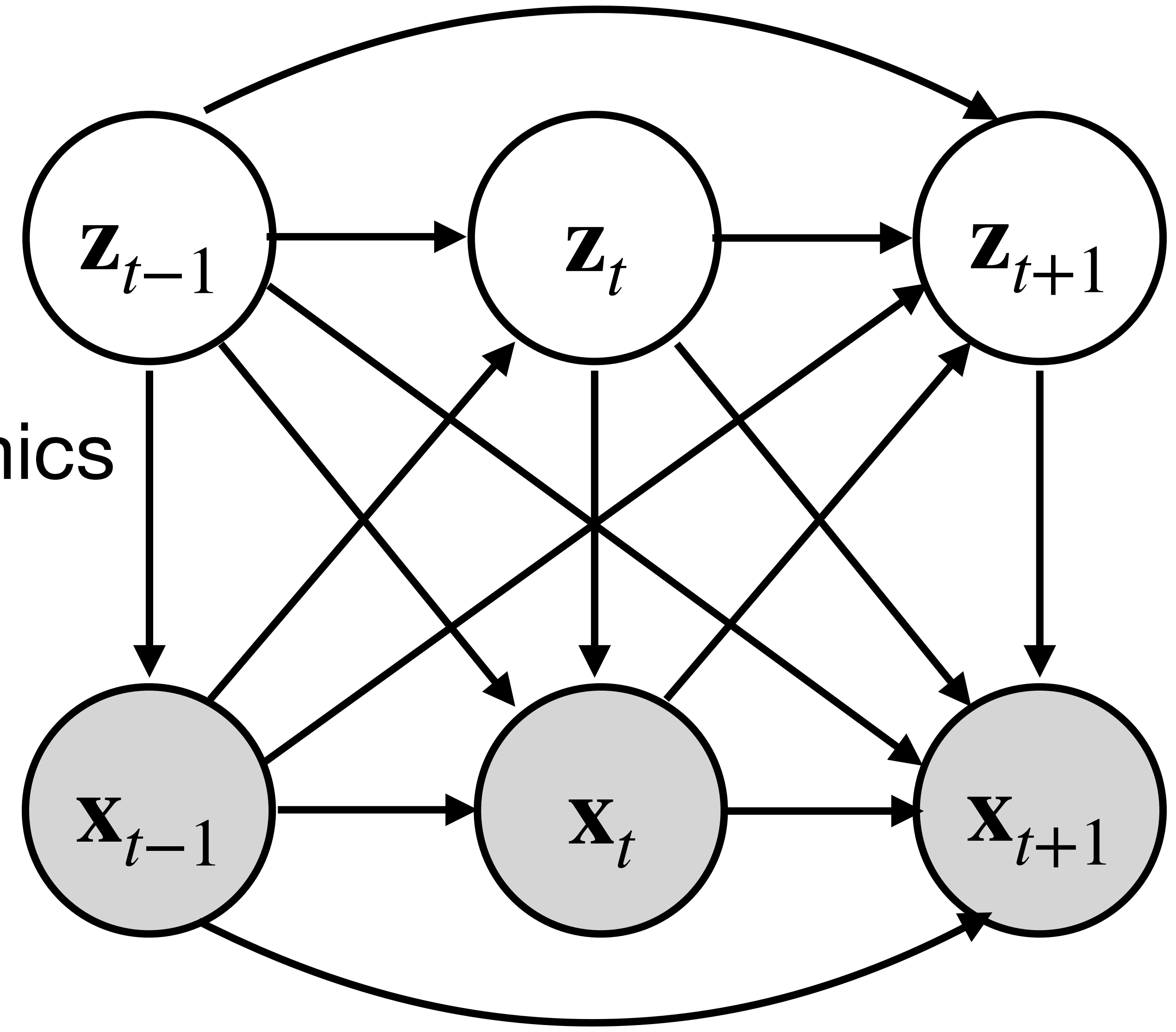


$p_\theta(x) = \mathcal{N}(\mu, \sigma^2)$ $\quad$ $p_\theta(x) = \int p_\theta(x \mid z) p_\theta(z) dz$

Example: probabilistic sequential data models



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)

$\mathbf{z}$ discrete

$\mathbf{z}$ continuous and Linear dynamics

Hidden Markov Model (HMM)

Linear Dynamical System (LDS)

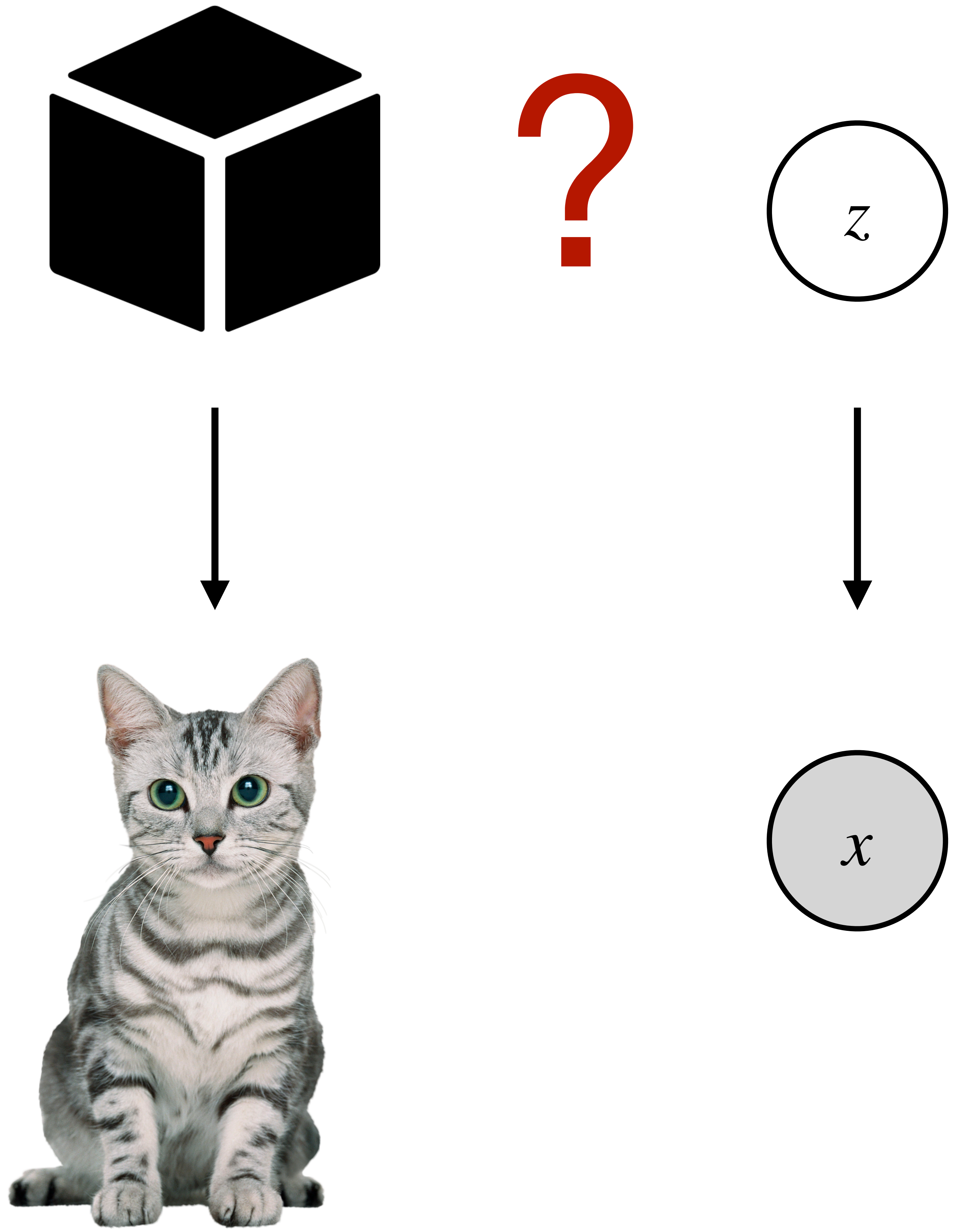
Dynamical Variational Auto-encoders (DVAEs)

Two main objectives of latent variable models

- Help to construct more complex distributions

$$p_\theta(x) = \int p_\theta(x \mid z) p_\theta(z) dz$$

- Infer the unknown variables



Two main objectives of latent variable models

- Help to construct more complex distributions

$$p_\theta(x) = \int p_\theta(x \mid z) p_\theta(z) dz$$

- Infer the unknown variables : Bayesian Inference

likelihood

prior

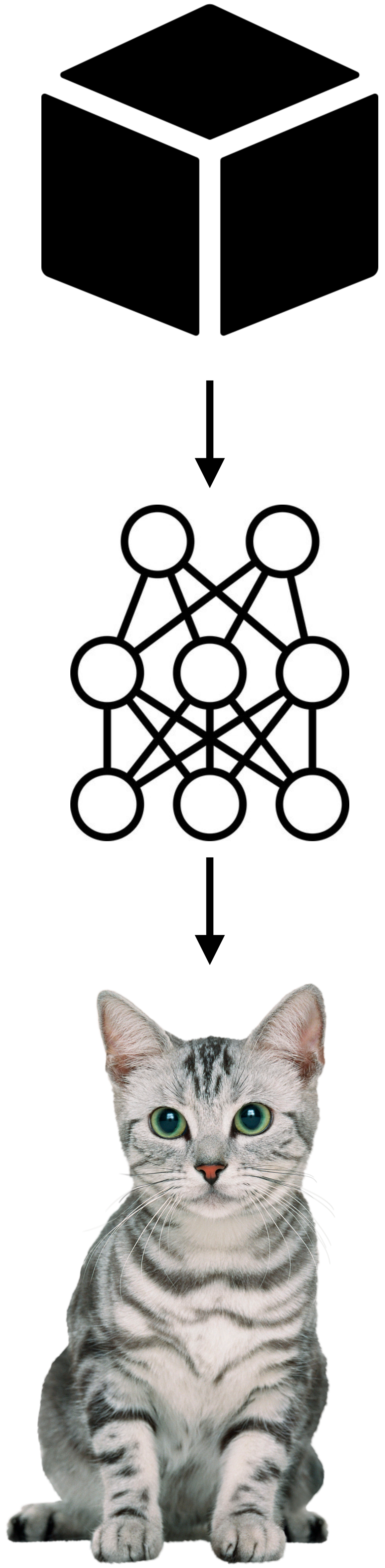
$$p_\theta(z \mid x) = \frac{\boxed{p_\theta(x \mid z)}\,\boxed{p_\theta(z)}}{\boxed{\int p_\theta(x \mid z) p_\theta(z) dz}}$$

posterior

Intractable!

marginal likelihood / evidence

$p_\theta(x \mid z)$



Variational Inference (VI)

- Solution: introduce a variational distribution to approximate the posterior

$$q(z) \approx p_\theta(z \mid x)$$

Variational Inference (VI)

- Solution: introduce a variational distribution to approximate the posterior

$$q(z) \approx p_\theta(z|x)$$

- Optimisation: minimize the Kullback-Leibler (KL) divergence

$$KL[q(z)||p_\theta(z|x)] = -\mathbb{E}_{q(z)}[\log \frac{p_\theta(z|x)}{q(z)}]$$

Intractable

Variational Inference (VI)

- Solution: introduce a variational distribution to approximate the posterior

$$q(z) \approx p_\theta(z \mid x)$$

- Optimisation: minimize the Kullback-Leibler (KL) divergence

$$KL[q(z) \mid\mid p_\theta(z \mid x)] = -\mathbb{E}_{q(z)}[\log \frac{p_\theta(z \mid x)}{q(z)}]$$

$$= -\mathbb{E}_{q(z)}[\log \frac{p_\theta(x, z)}{p_\theta(x) q(z)}] = -\mathbb{E}_{q(z)}[\log \frac{p_\theta(x, z)}{q(z)} - \log p_\theta(x)]$$

Independent with $q(z)$

Variational Inference (VI)

- Solution: introduce a variational distribution to approximate the posterior

$$q(z) \approx p_\theta(z|x)$$

- Optimisation: minimize the Kullback-Leibler (KL) divergence

$$KL[q(z)||p_\theta(z|x)] = -\mathbb{E}_{q(z)}[\log \frac{p_\theta(z|x)}{q(z)}]$$

$$= -\mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{p_\theta(x)q(z)}] = -\mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{q(z)} - \log p_\theta(x)]$$

Model evidence

$$= \log p_\theta(x) - \mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{q(z)}]$$

Variational Inference (VI)

$$KL[q(z)\,||\,p_\theta(z\,|\,x)] = \log p_\theta(x) - \mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{q(z)}]$$

Model evidence

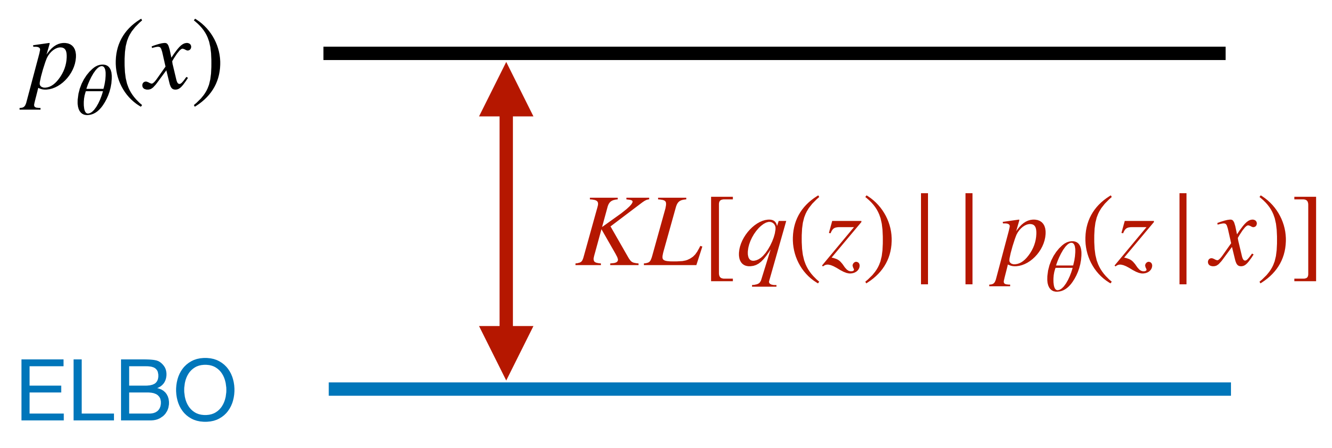
Minimize $KL[q(z)\,||\,p_\theta(z\,|\,x)]$ w.r.t $q(z)$ $\Leftrightarrow$ Maximize $\mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{q(z)}]$ w.r.t $q(z)$

$$\mathbb{E}_{q(z)}[\log \frac{p_\theta(x,z)}{q(z)}] = \log p_\theta(x) - KL[q(z)\,||\,p_\theta(z\,|\,x)] \leq \log p_\theta(x)$$

≥ 0

Evidence Lower BOund (ELBO)

Maximum log likelihood $\log p_\theta(x)$ $\Rightarrow$ Maximize ELBO

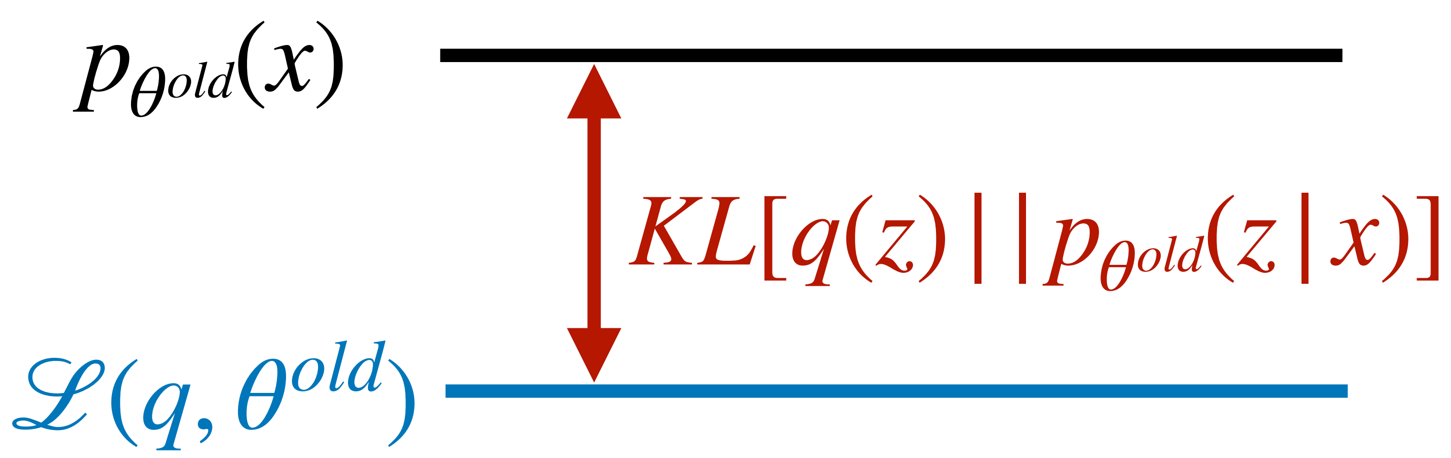


Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$

Variational Inference (VI)

- ELBO: $$\mathcal{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm



E step: set $q(z) = p_{\theta^{old}}(z\,|\,x)$ and compute $\mathcal{Q}(\theta,\theta^{old}) = \mathbb{E}_{p_{\theta^{old}}(z|x)}[\log p_\theta(x,z)]$

Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm

$\mathscr{L}(q,\theta^{old})$ $p_{\theta^{old}}(x)$ $KL[q(z)\,||\,p_{\theta^{old}}(z\,|\,x)] = 0$

E step: set $q(z) = p_{\theta^{old}}(z\,|\,x)$ and compute $\mathscr{Q}(\theta,\theta^{old}) = \mathbb{E}_{p_{\theta^{old}}(z|x)}[\log p_\theta(x,z)]$

Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm

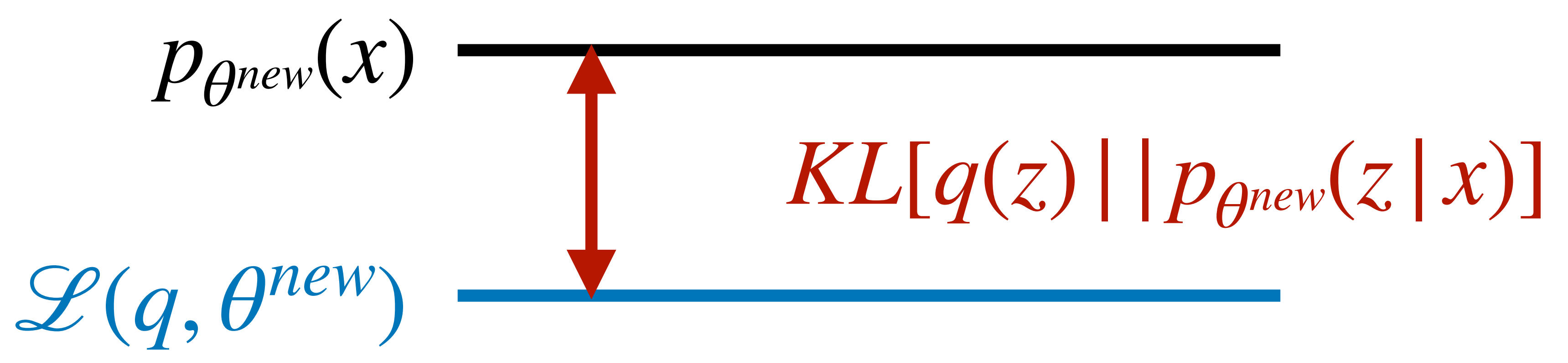
$\mathscr{L}(q, \theta^{old})$ $p_{\theta^{old}}(x)$ $KL[q(z)||p_{\theta^{old}}(z|x)] = 0$

M step: estimate $\theta^{new} = \arg\max_\theta \mathscr{Q}(\theta, \theta^{old})$

Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$

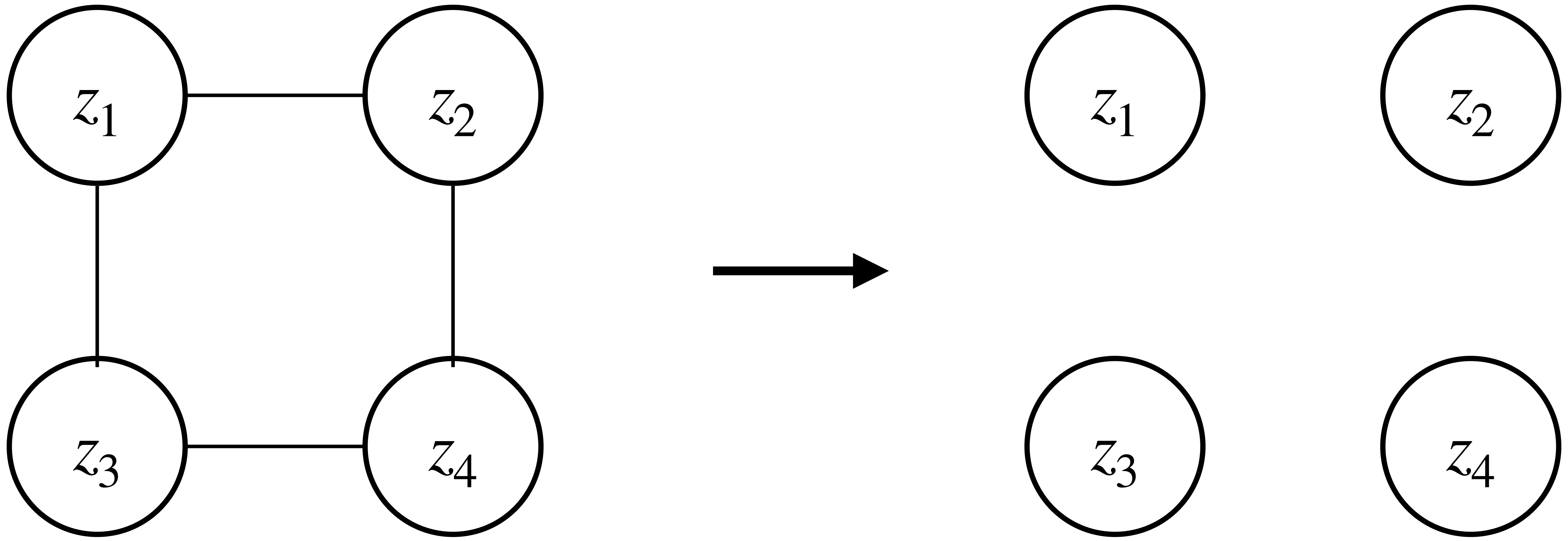
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm



M step: estimate $\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old})$

Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm
- Mean-field approximation: $$q(z) = \prod_{i=1}^{M} q_i(z_i \mid x)$$



Variational Inference (VI)

- ELBO: $$\mathscr{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm
- Mean-field approximation: $q(z) = \prod_{i=1}^{M} q_i(z_i \mid x)$ $\longrightarrow$ Variational EM algorithm

Variational E Step: $\forall i \in \{1,\ldots M\},$ compute $q_i(z_i \mid x) \propto \exp\left(\mathbb{E}_{\prod_{j\neq i}}[q_j(z_j \mid x)]\right)$

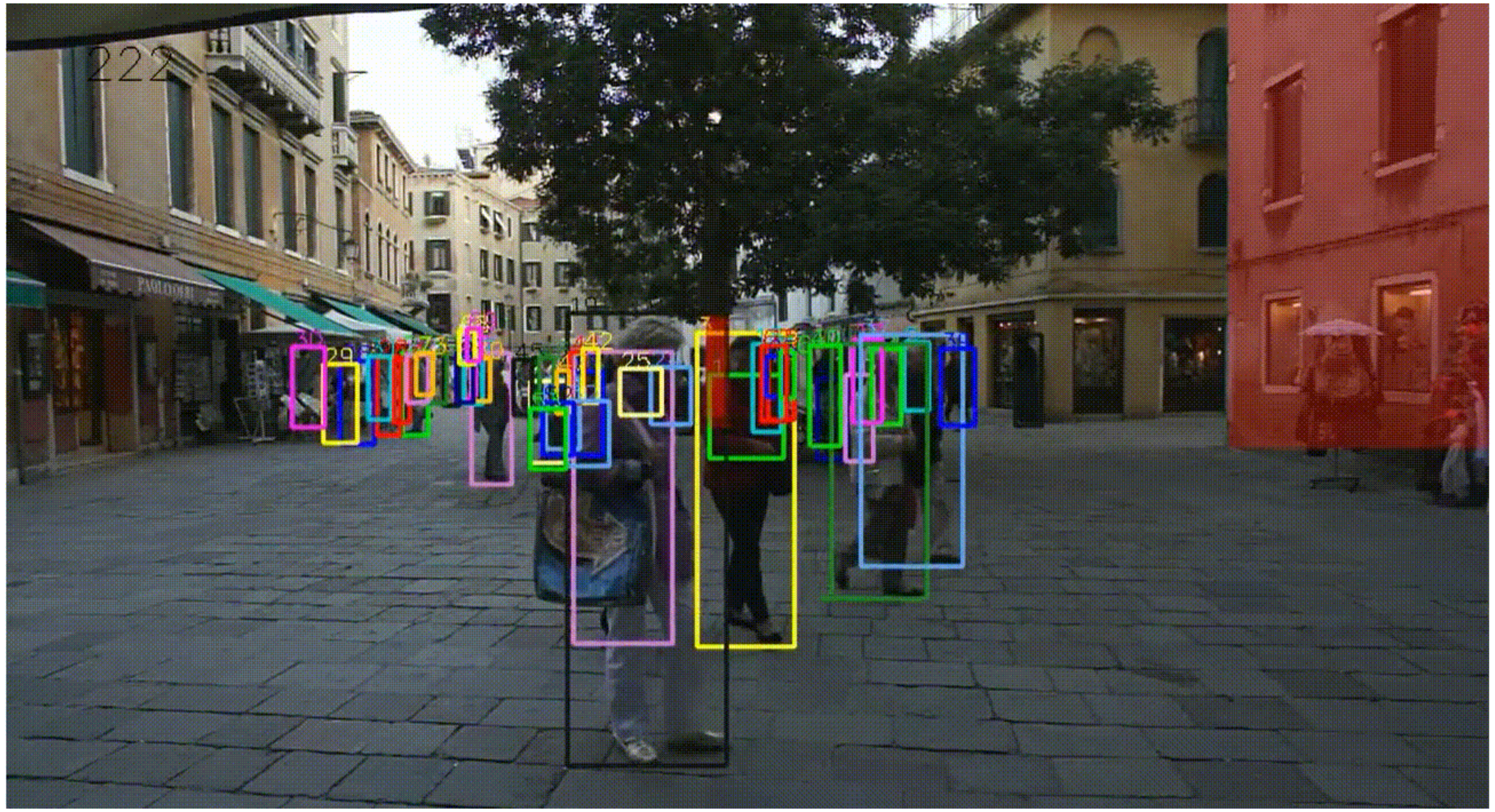
M Step: estimate $\theta^{new} = \arg\max_\theta \mathscr{L}(q,\theta)$

Variational Inference (VI)

- ELBO: $$\mathcal{L}(q,\theta) = \mathbb{E}_{q(z)}[\log p_\theta(x,z) - \log q(z)]$$
- If $q(z)$ can be expressed in closed form $\longrightarrow$ EM algorithm
- Mean-field approximation: $q(z) = \prod_{i=1}^{M} q_i(z_i \mid x)$ $\longrightarrow$ Variational EM algorithm
- Amortized inference: $\mathcal{L}(\phi,\theta) = \mathbb{E}_{q_\phi(z)}[\log p_\theta(x \mid z)] - KL\left(q_\phi(z) \,||\, p(z)\right)$ $\longrightarrow$ VAE

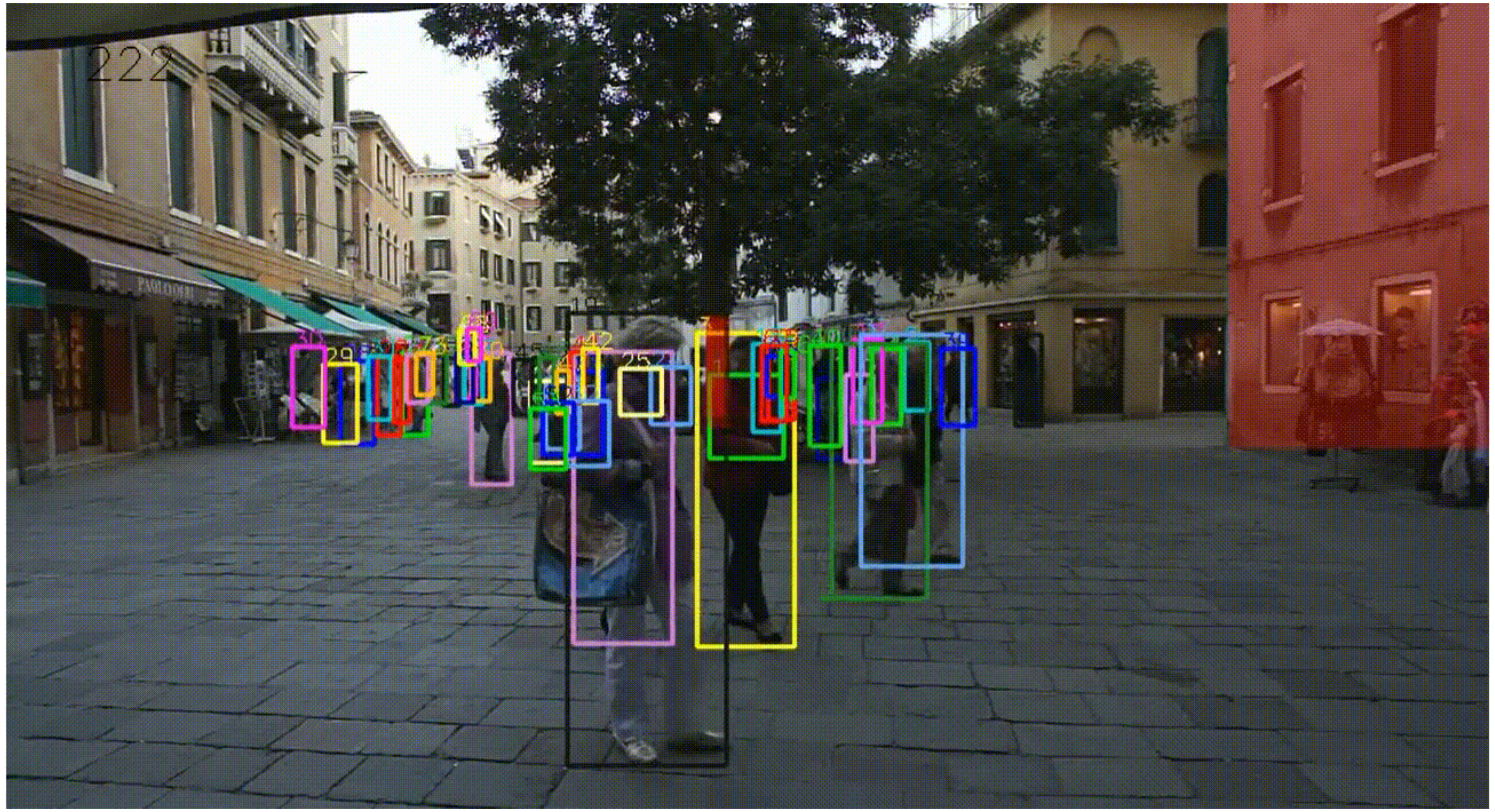
Unsupervised multi-object tracking (MOT) with MixDVAE

MOT task definition



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

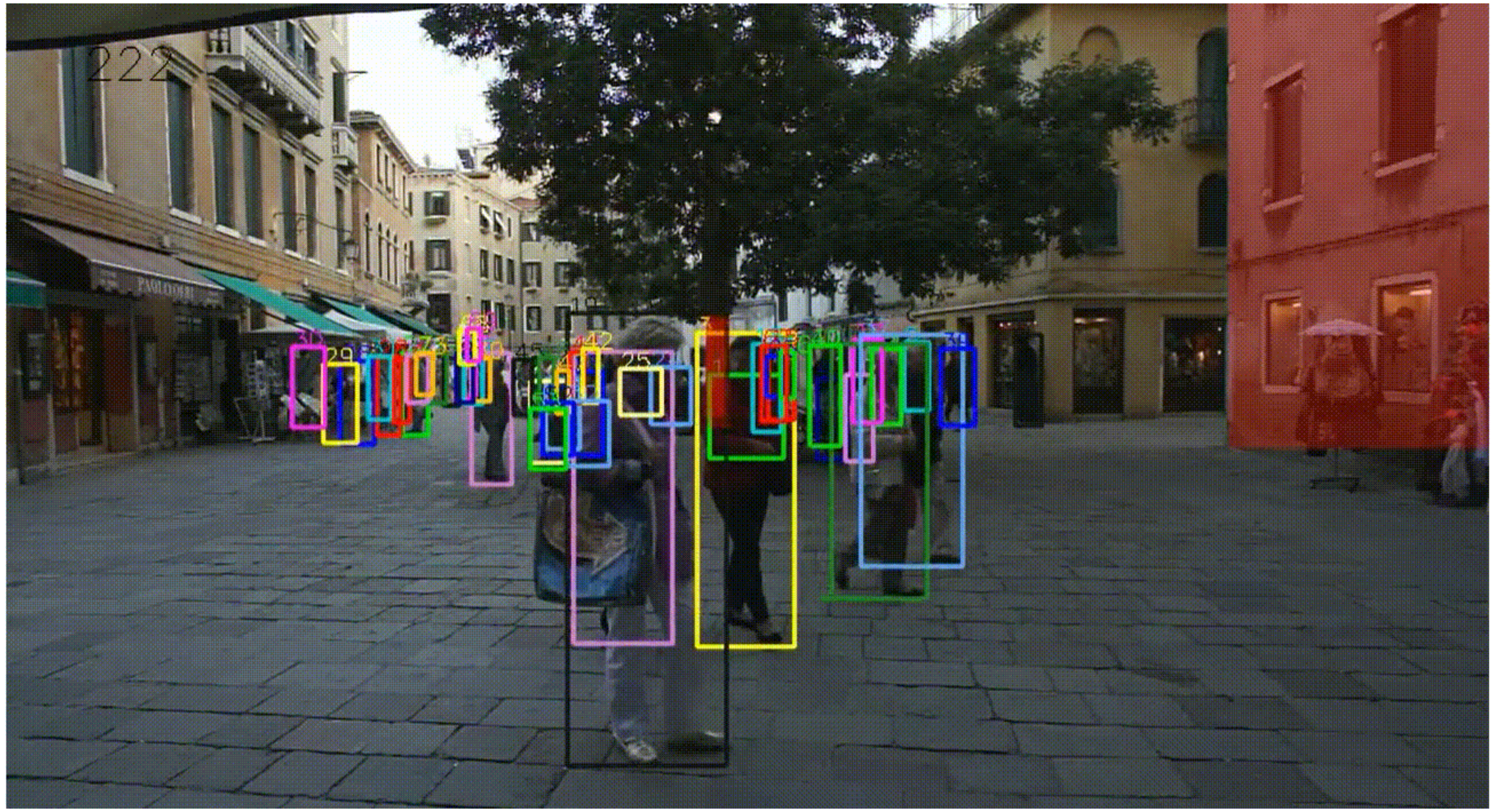
MOT task definition



4 main sub-tasks in MOT

- Extracting source observations (detections) at each time frame

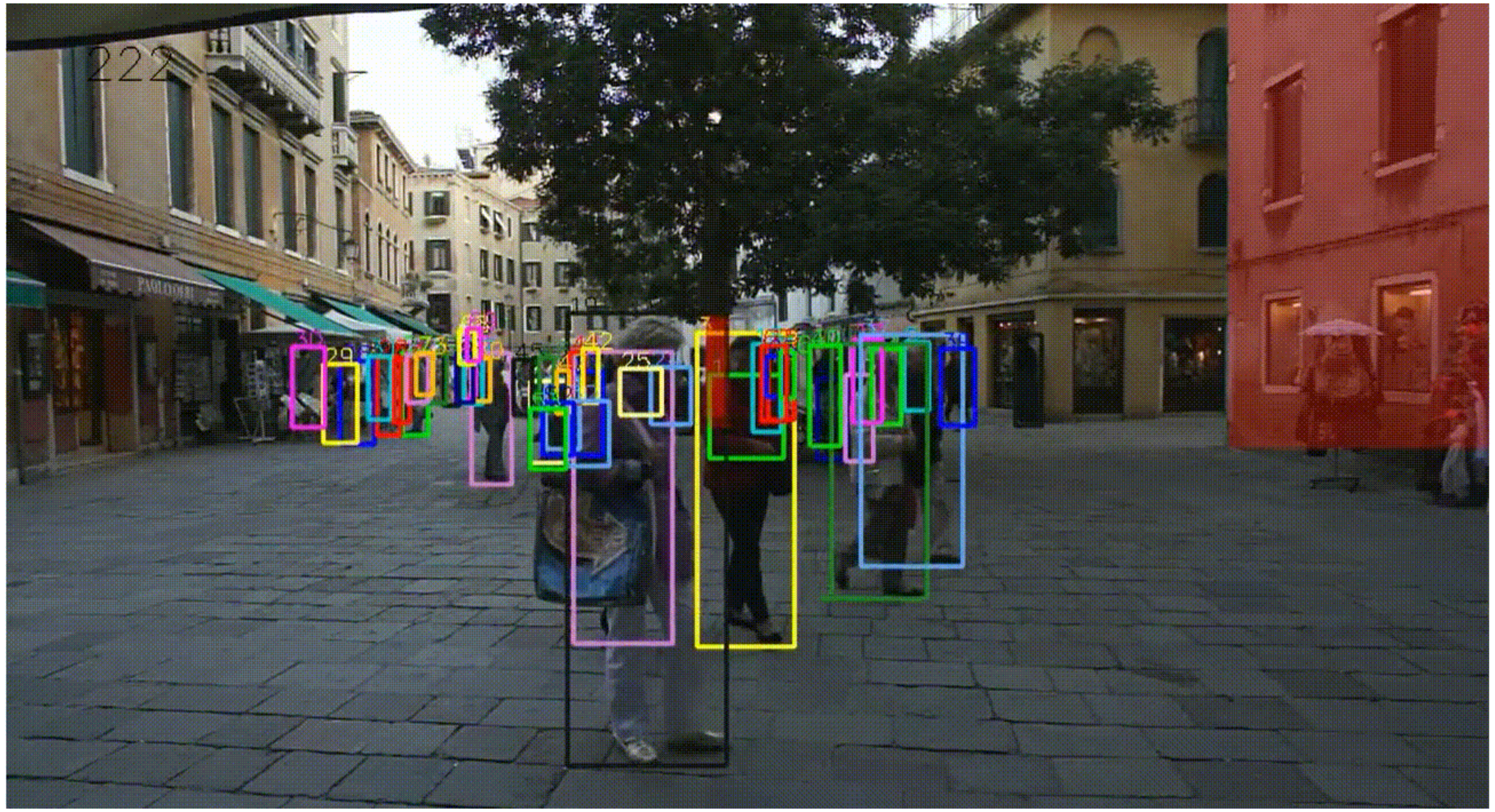
MOT task definition



4 main sub-tasks in MOT

- Extracting source observations (detections) at each time frame
- Modeling the dynamics of the sources' movements

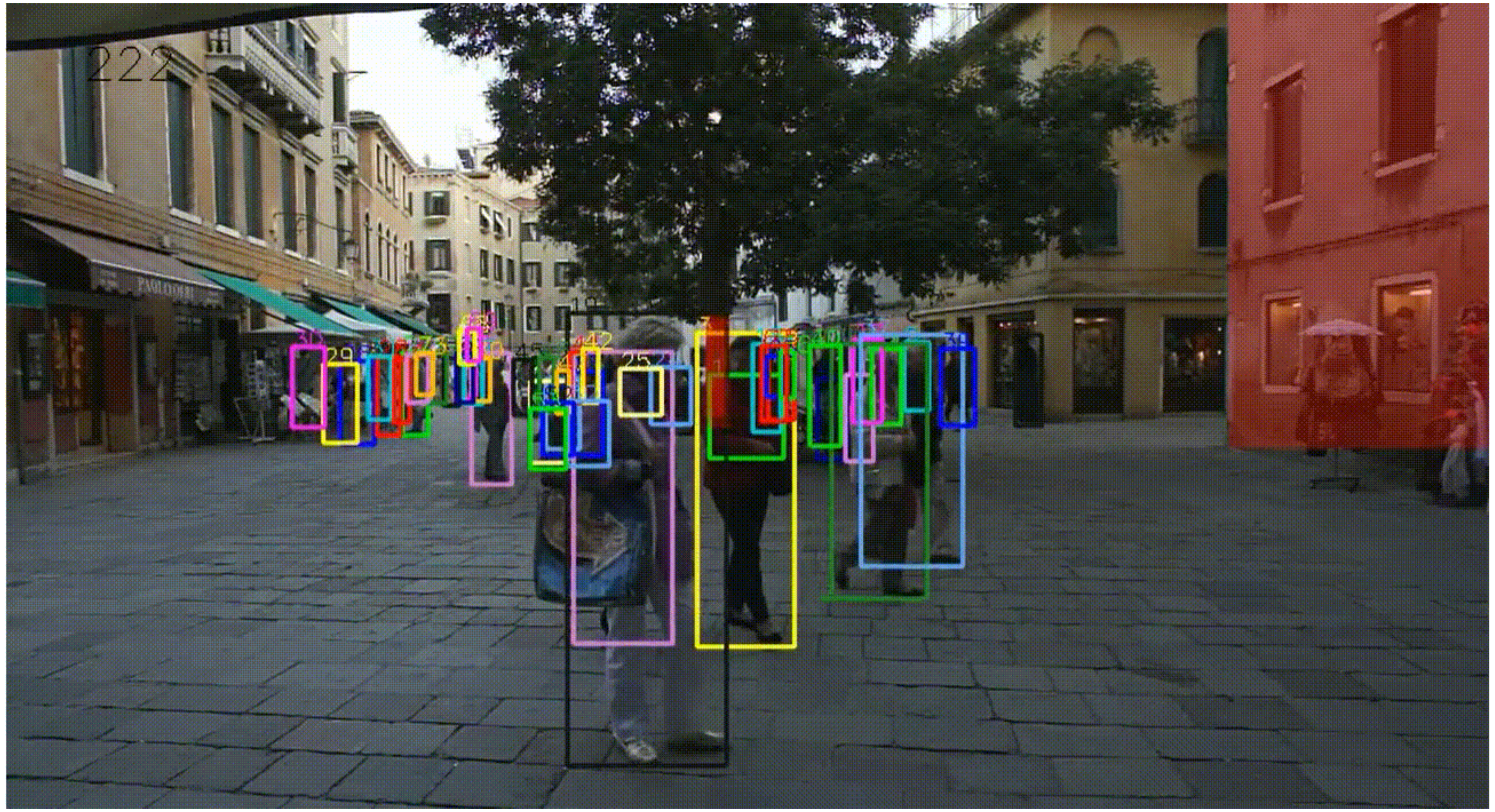
MOT task definition



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

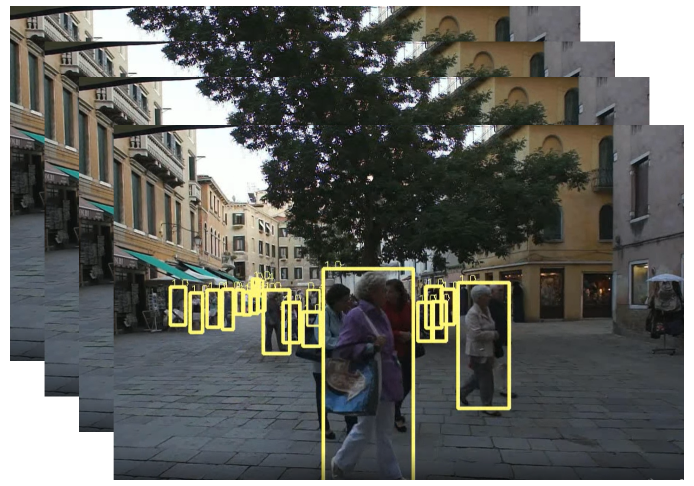
MOT task definition



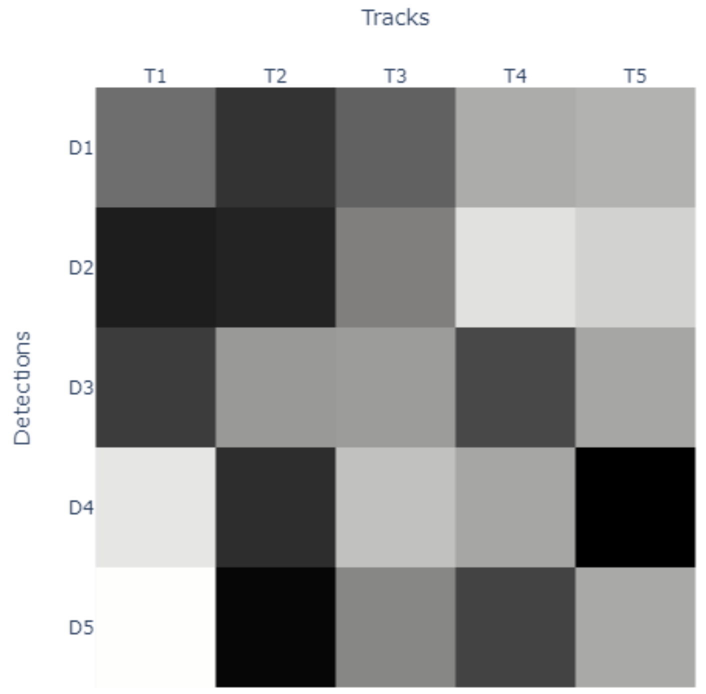
4 main sub-tasks in MOT

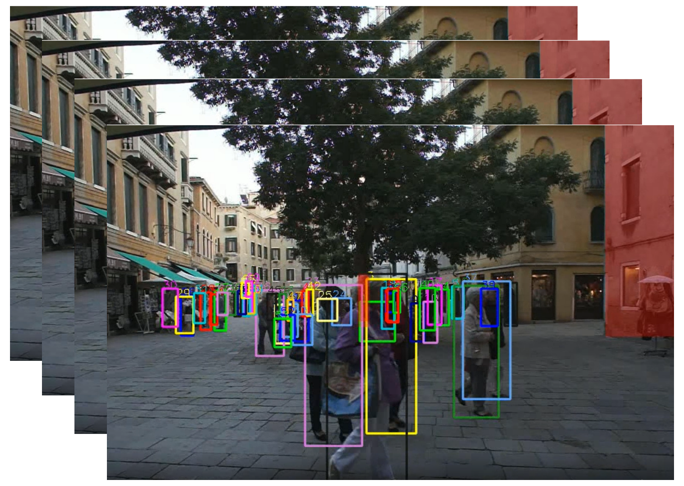
- Extracting source observations (detections) at each time frame
- Modeling the dynamics of the sources' movements
- Associating observations to sources consistently over time
- Accounting for birth and death process of source trajectories

Tracking-by-Detection paradigm



- Appearance
- Motion
- Interactions
- …





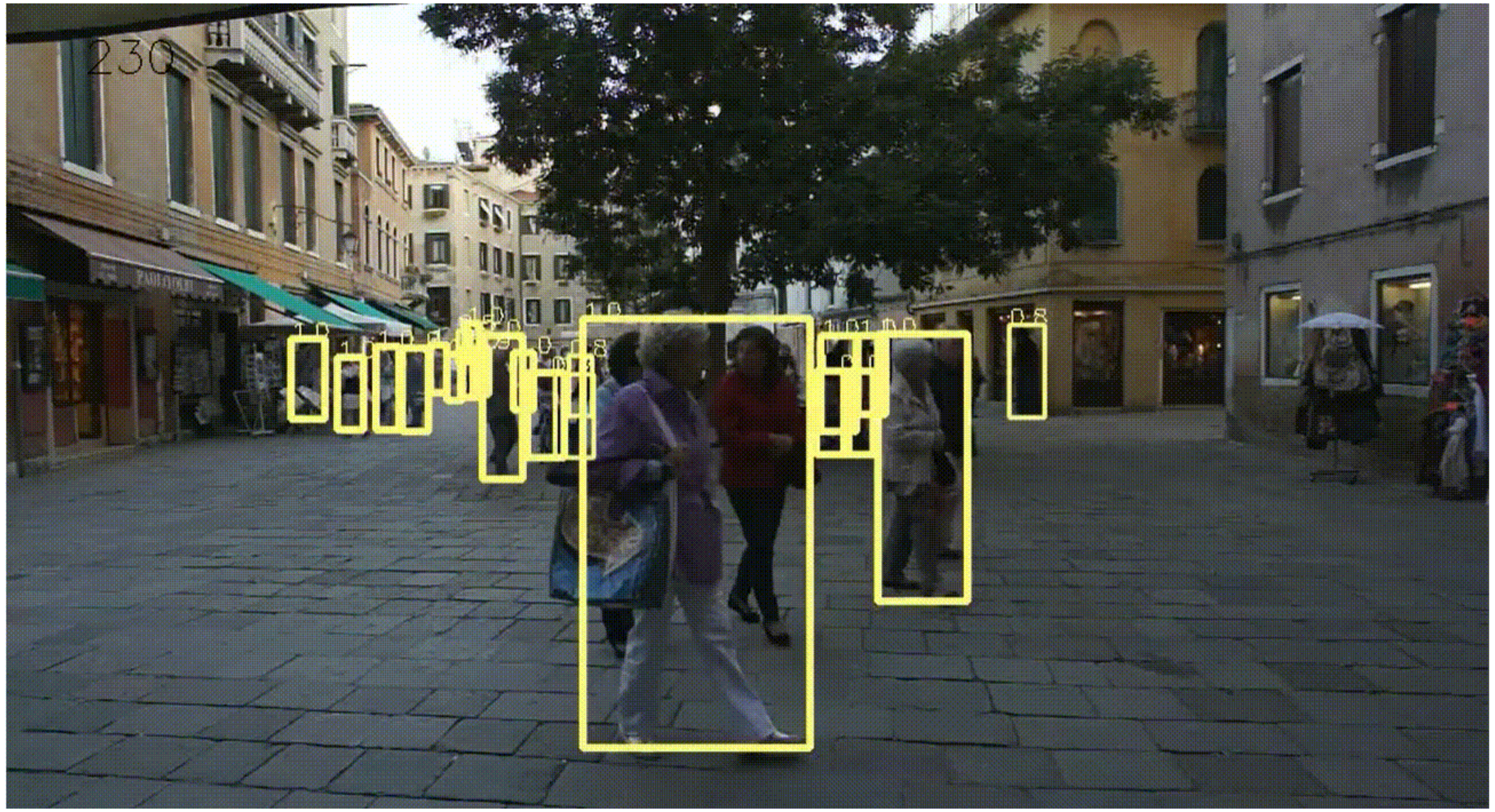
Detection

Feature Extraction

Similarity / Cost Calculation

Data association, Re-Identification, Trajectories birth/death

Challenges



Detections by a public detector SDP (Yang et al., 2016)

Appearance related issues

- Camera motion
- Bad illumination
- Objects occlusion
- Similar appearances
- Noisy detections

ID Switches
False Negatives

Modelling sources' motion dynamics

Motion models

- Constant velocity assumptions / Kalman Filter (Bewley et al., 2016; Woke et al., 2017; Bergmann et al., 2019; Ban et al., 2021)
- RNN / Neural Network based models (Milan et al., 2017; Sadeghian et al., 2017; Babaee et al., 2018)
- Probabilistic motion models (Fang et al., 2018; Saleh et al., 2021)

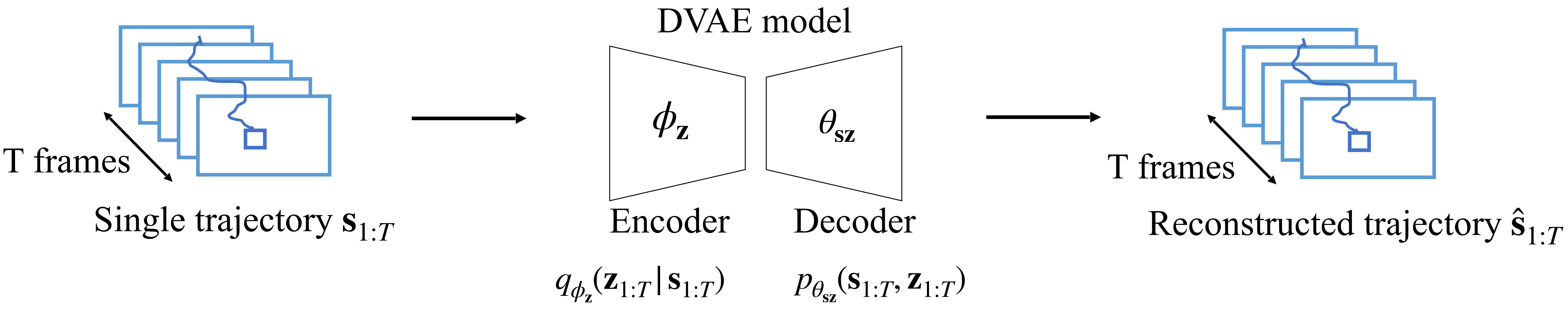
Challenges

low video sampling rate, moving camera, high object velocity, complex non-linear motion patterns in long-term tracks

NEEDS FOR MORE ROBUST MOTION MODELS

Use DVAEs for source motion dynamics modeling

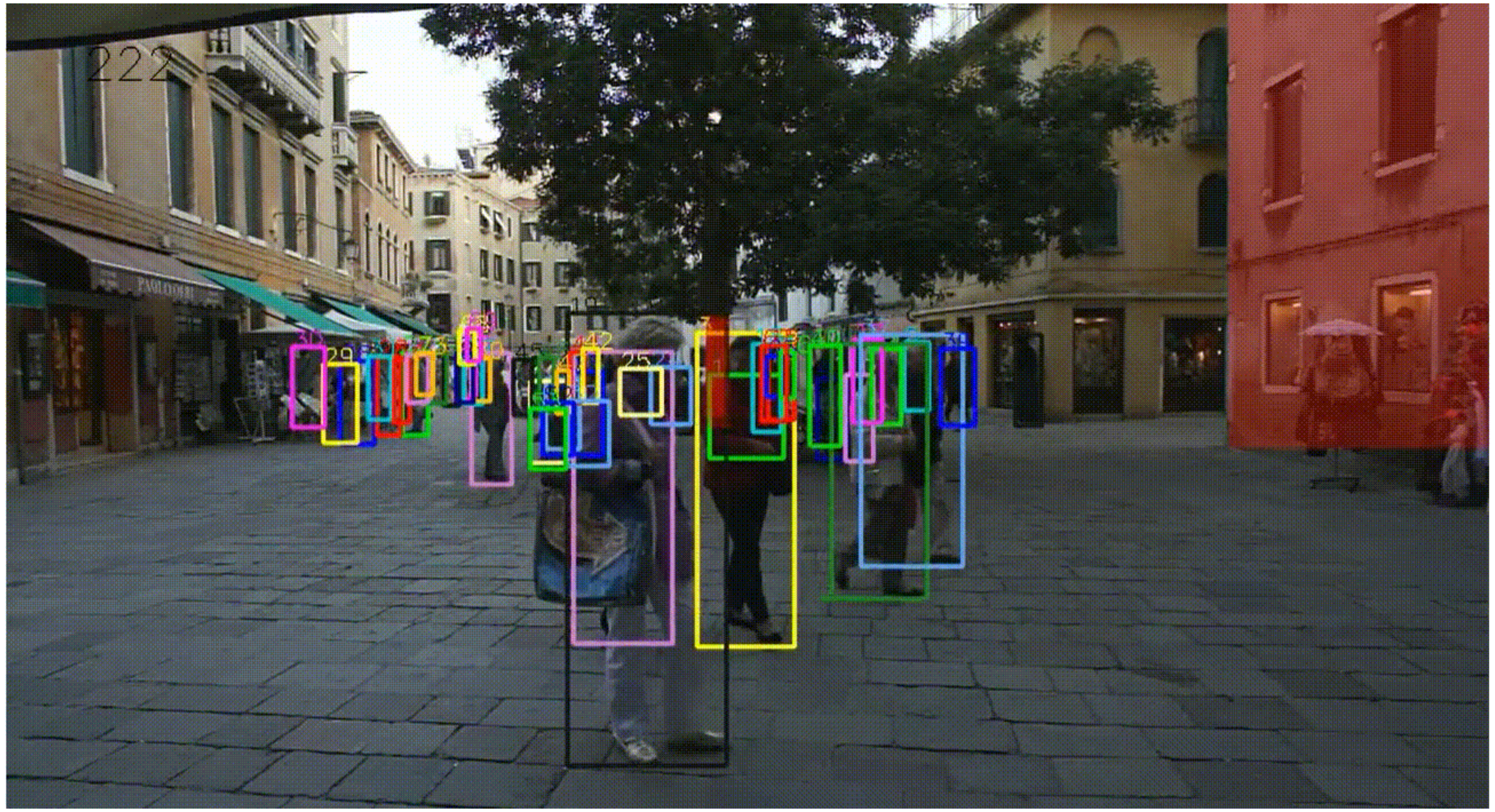
Non-linear probabilistic sequential latent variable generative models



Training by maximizing the Evidence Lower BOund (ELBO)

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

Focus on two sub-tasks



4 main sub-tasks in MOT

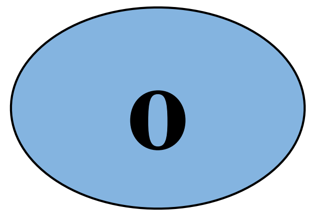
- Extracting source observations (detections) at each time frame
- Modeling the dynamics of the sources' movements
- Associating observations to sources consistently over time
- Accounting for birth and death process of source trajectories

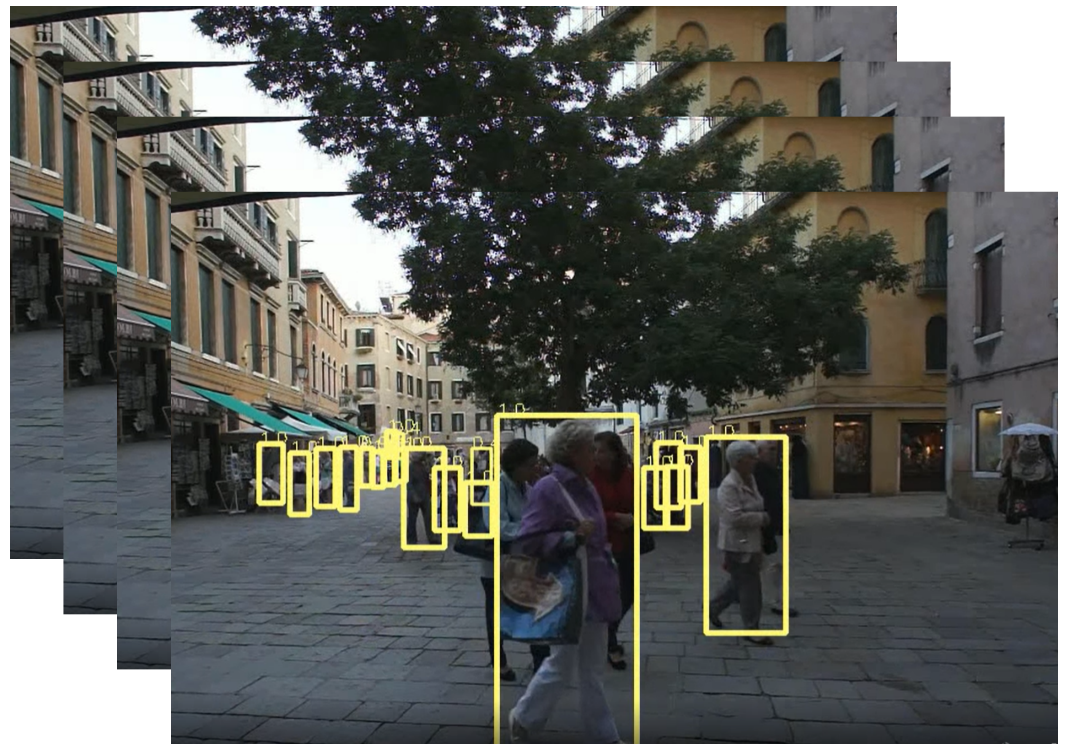
→ Tracking-by-detection, kown number of sources

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

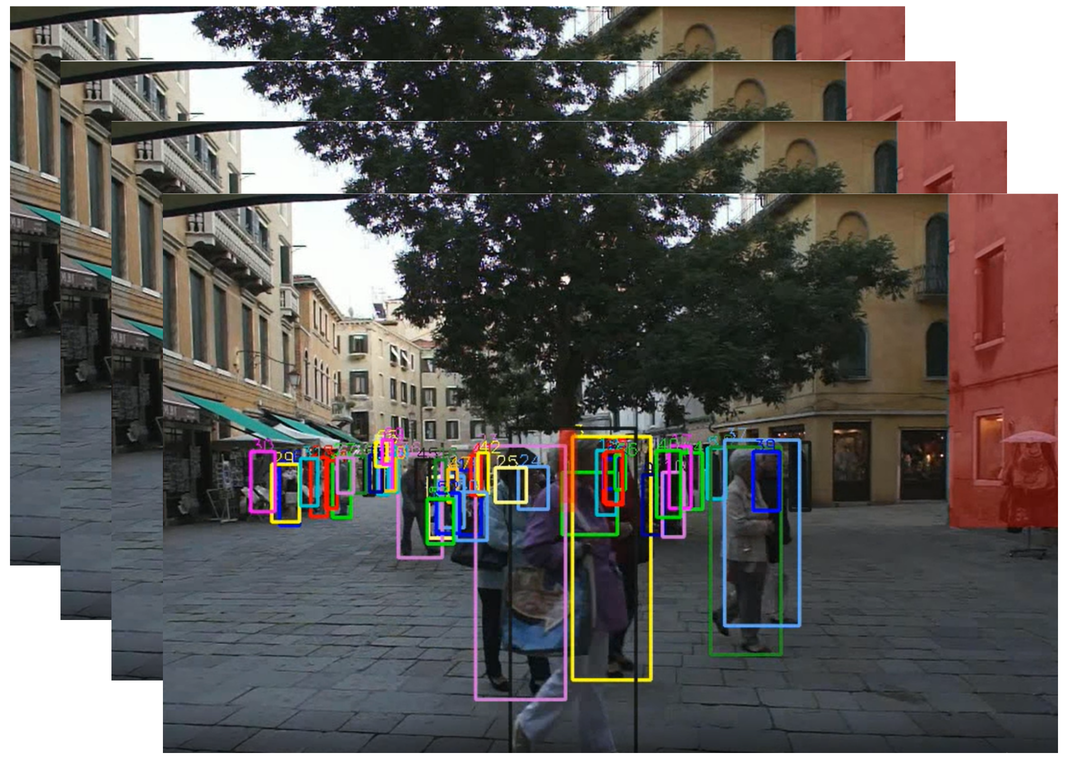


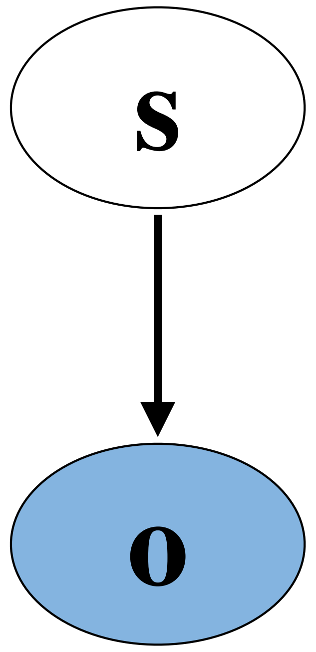


Define MOT from a probabilistic perspective

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

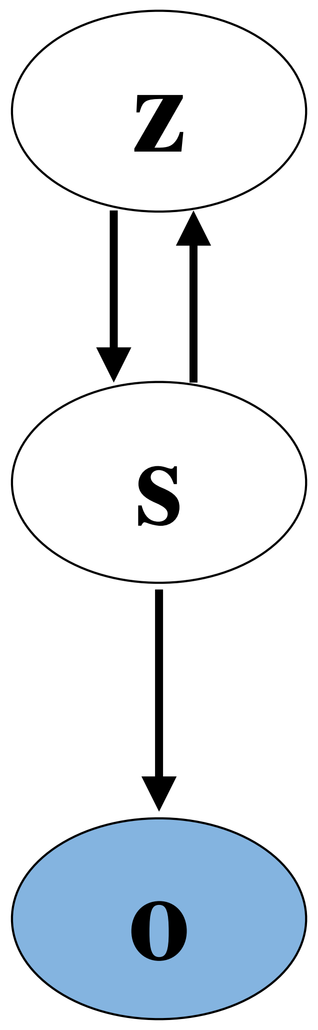




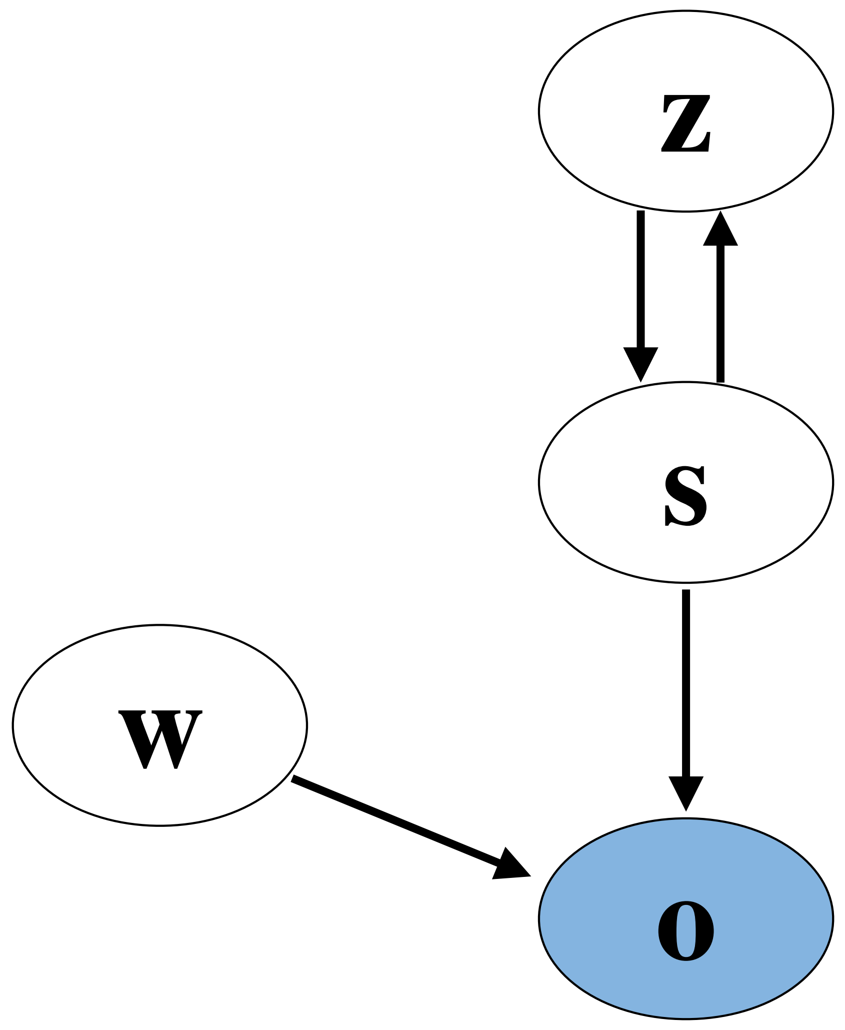
Define MOT from a probabilistic perspective

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



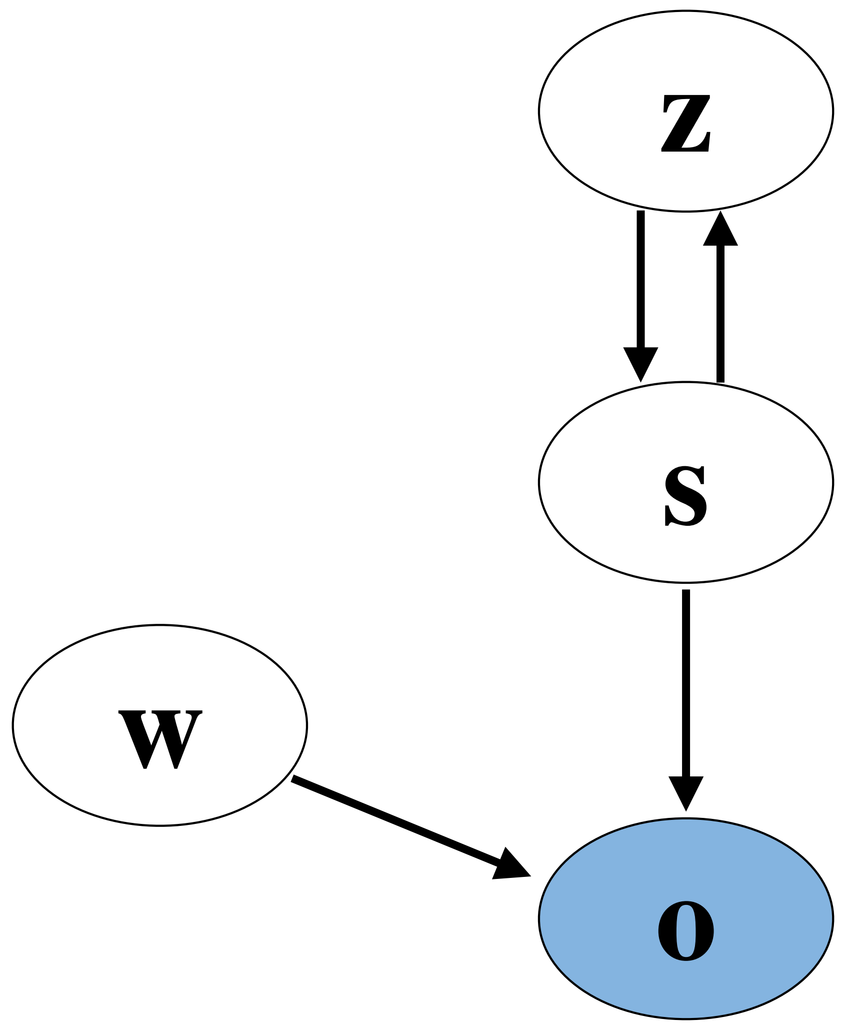
Define MOT from a probabilistic perspective



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

Define MOT from a probabilistic perspective



Definition of random variables

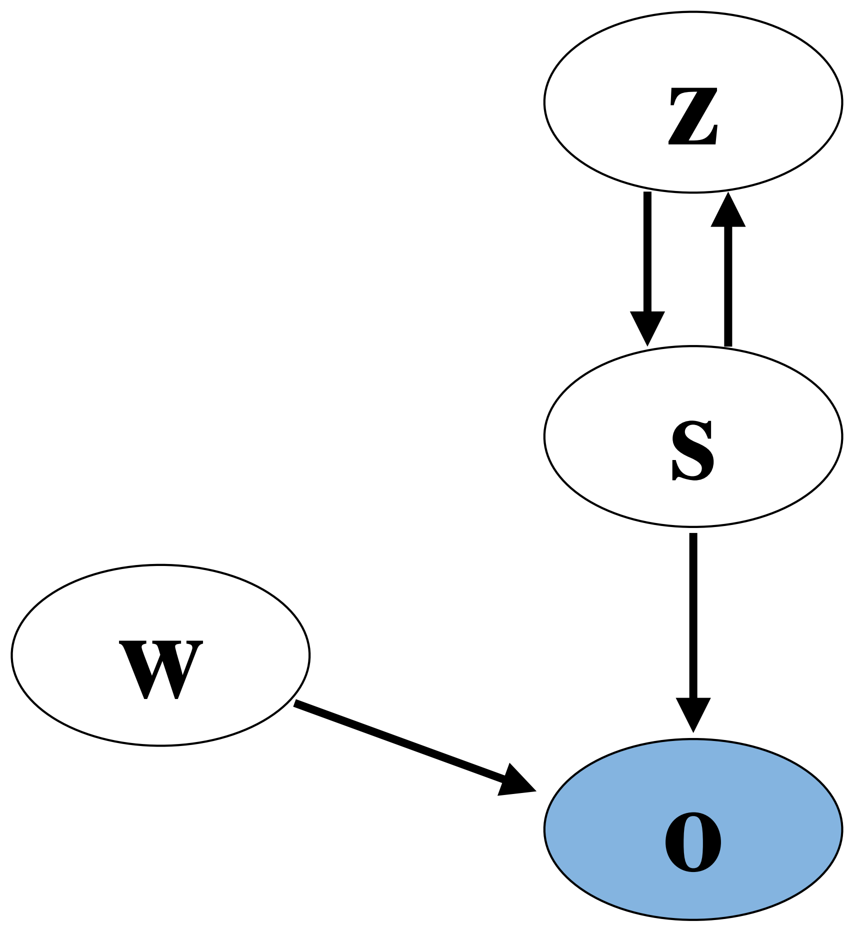
- $\mathbf{o} = \{\mathbf{o}_{1:T,1:K_t}\} \in \mathbb{R}^{T \times K_t \times 4}$: positions of detection bounding boxes
- $\mathbf{s} = \{\mathbf{s}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times 4}$: true positions of sources
- $\mathbf{z} = \{\mathbf{z}_{1:T,1:N}\} \in \mathbb{R}^{T \times N \times L}$: 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: $\mathbf{s}$, $\mathbf{z}$, $\mathbf{w}$

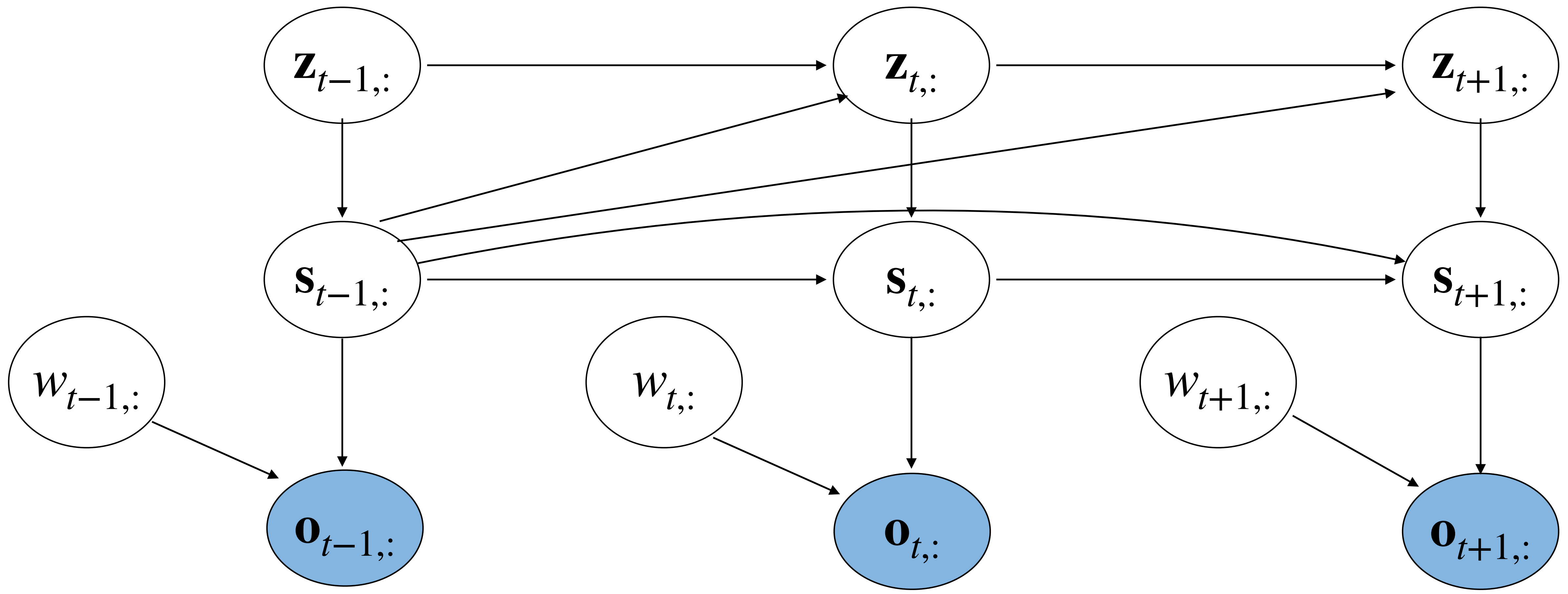
MOT objective: estimate the posterior distribution $p(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

Resolve MOT through Variational Inference (VI)

Associated graphical model



Folded graphical model

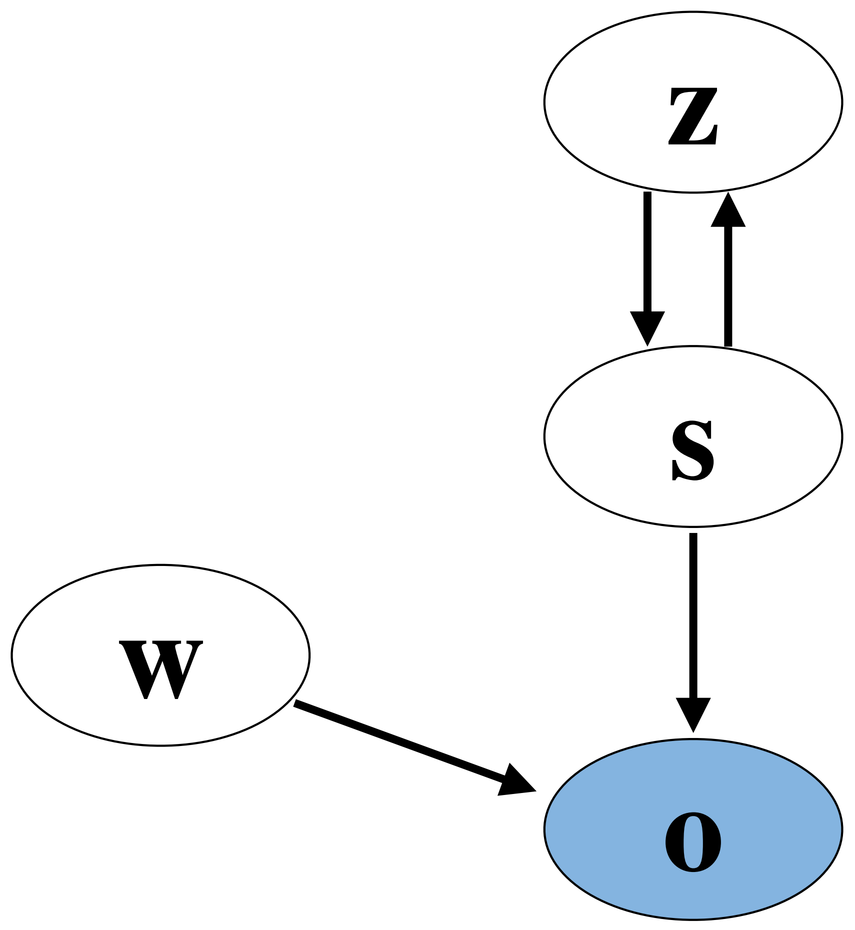


Extended graphical model over time frames

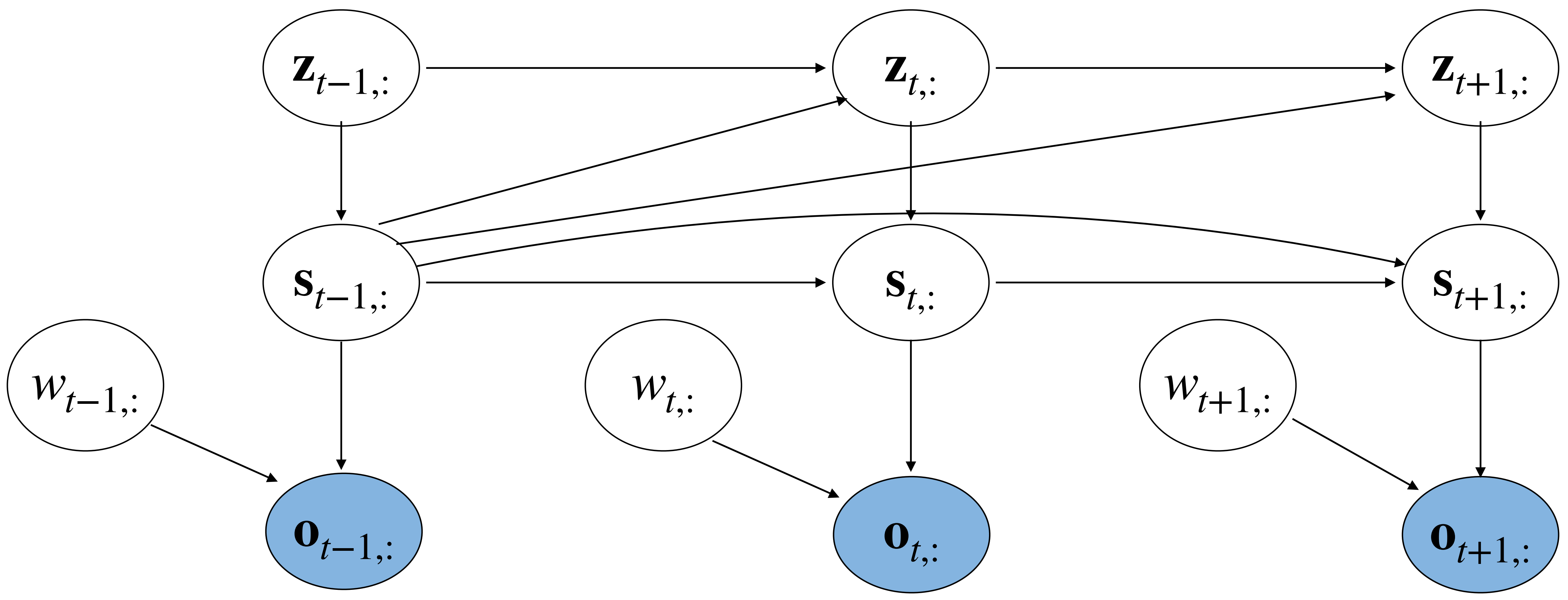
Generative model: $p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o} \mid \mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_{\mathbf{sz}}}(\mathbf{s}, \mathbf{z})$

Resolve MOT through Variational Inference (VI)

Associated graphical model



Folded graphical model



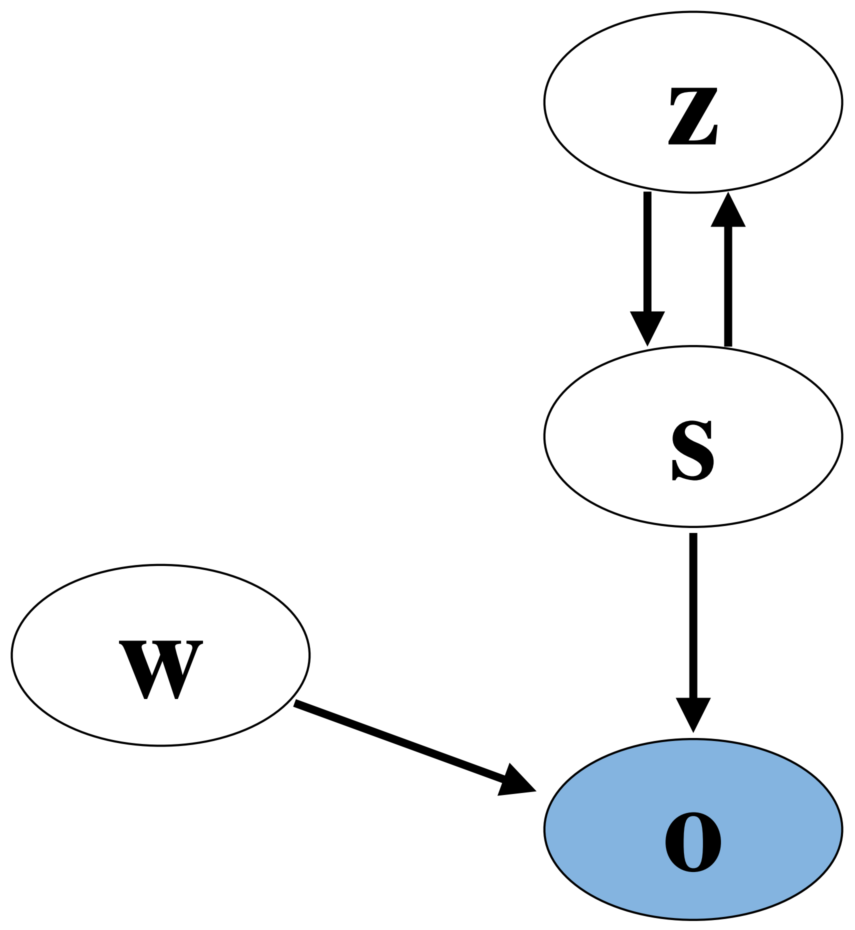
Extended graphical model over time frames

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

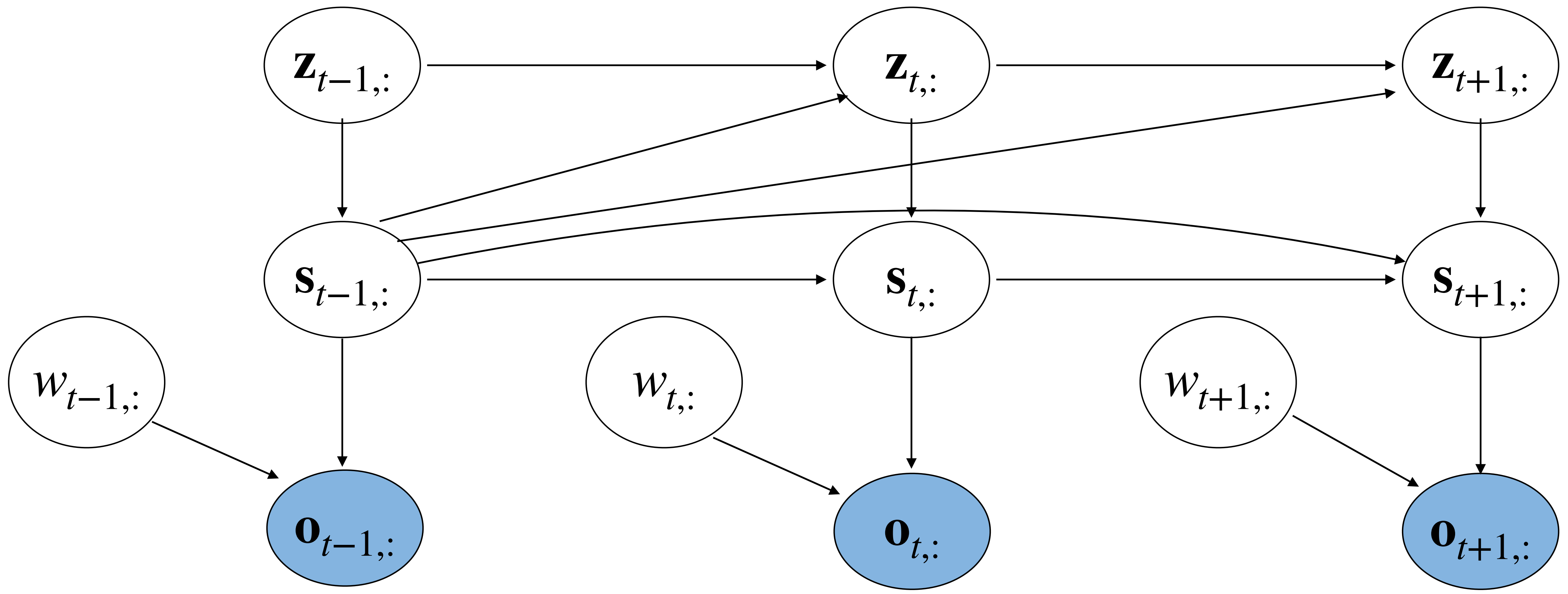
Intractable true posterior distribution $p_{\theta_{\mathbf{szw}}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

Resolve MOT through Variational Inference (VI)

Associated graphical model



Folded graphical model



Extended graphical model over time frames

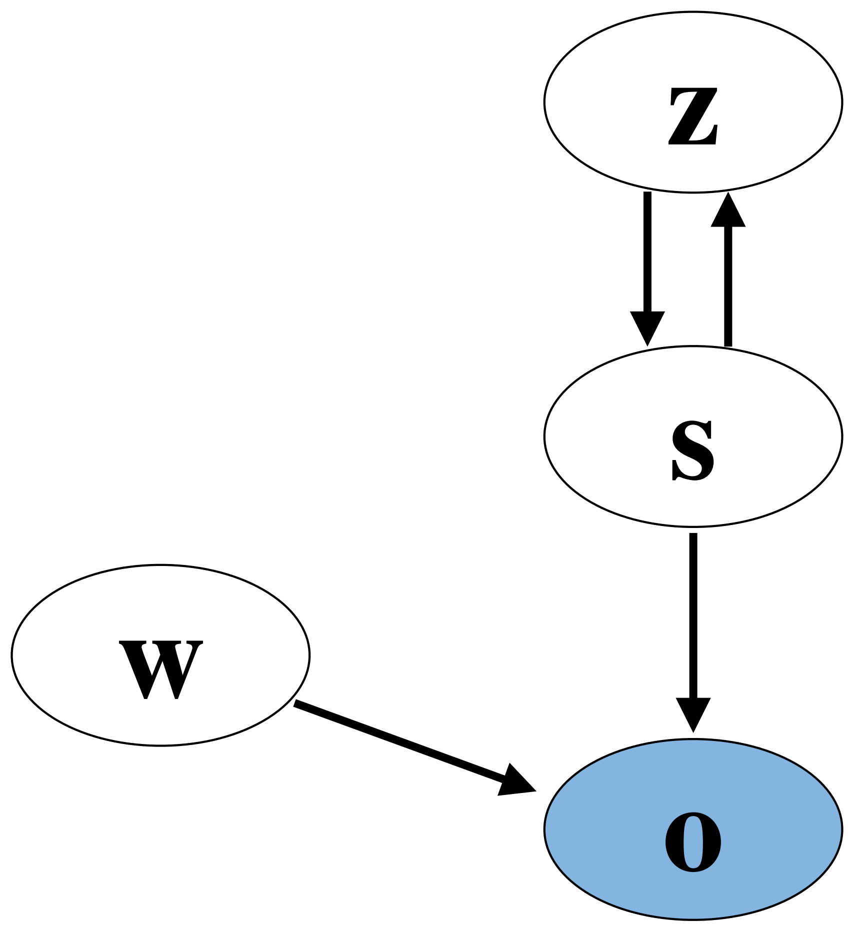
Generative model: $p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o} \mid \mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_\mathbf{sz}}(\mathbf{s}, \mathbf{z})$

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

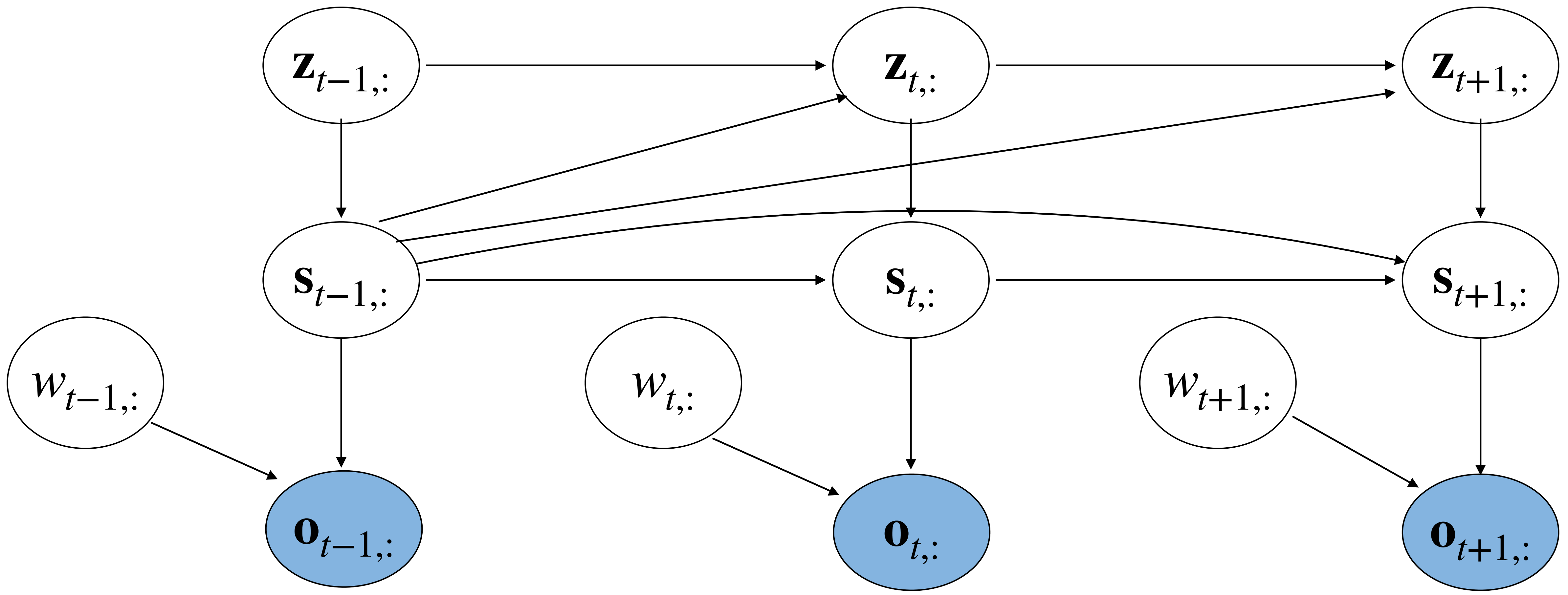
Inference model: mean-field like approximation $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o}) \approx q_{\phi_\mathbf{w}}(\mathbf{w} \mid \mathbf{o}) q_{\phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{s}) q_{\phi_\mathbf{s}}(\mathbf{s} \mid \mathbf{o})$

Resolve MOT through Variational Inference (VI)

Associated graphical model



Folded graphical model



Extended graphical model over time frames

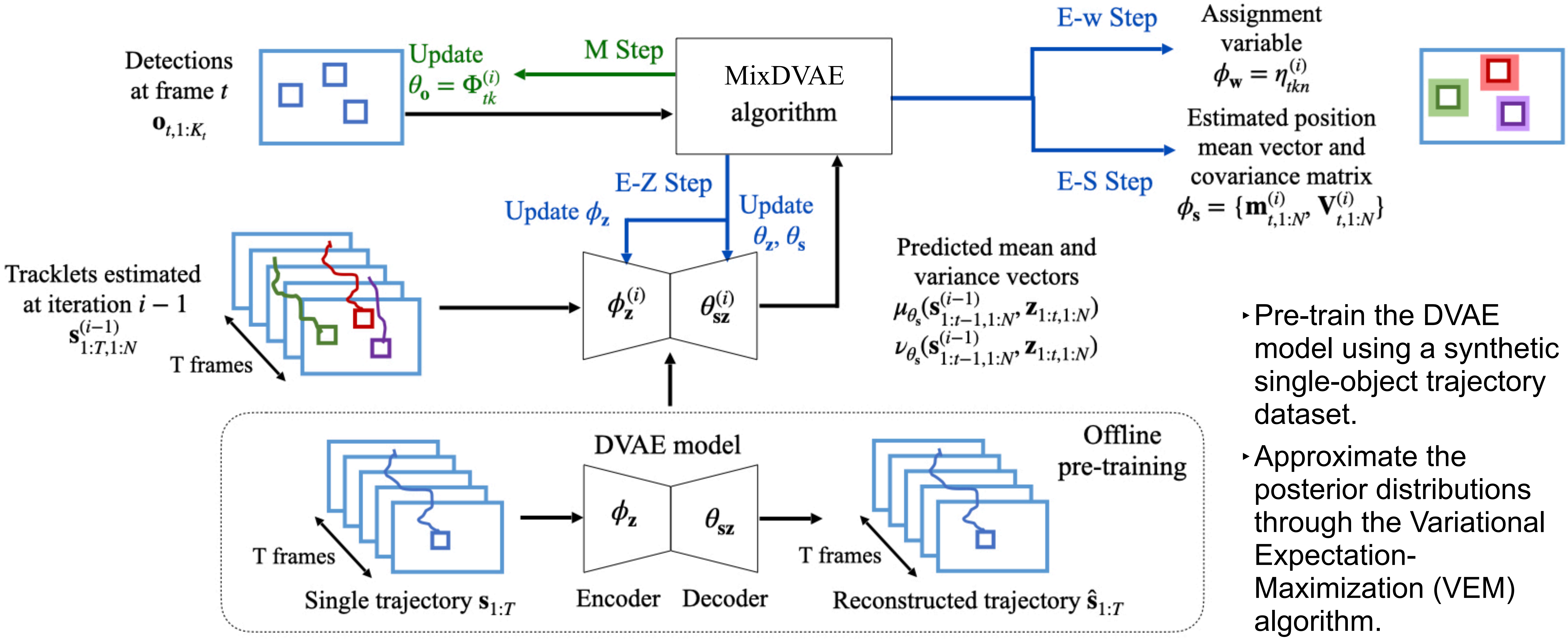
Generative model: $p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o} \,|\, \mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_\mathbf{sz}}(\mathbf{s}, \mathbf{z})$

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 $\mathcal{L}(\theta, \phi; \mathbf{o}) = \mathbb{E}_{q_\phi(\mathbf{s},\mathbf{z},\mathbf{w}|\mathbf{o})}[\log p_\theta(\mathbf{o}, \mathbf{s}, \mathbf{z}, \mathbf{w}) - \log q_\phi(\mathbf{s}, \mathbf{z}, \mathbf{w} \,|\, \mathbf{o})]$

Resolve MOT through Variational Inference (VI)



- 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

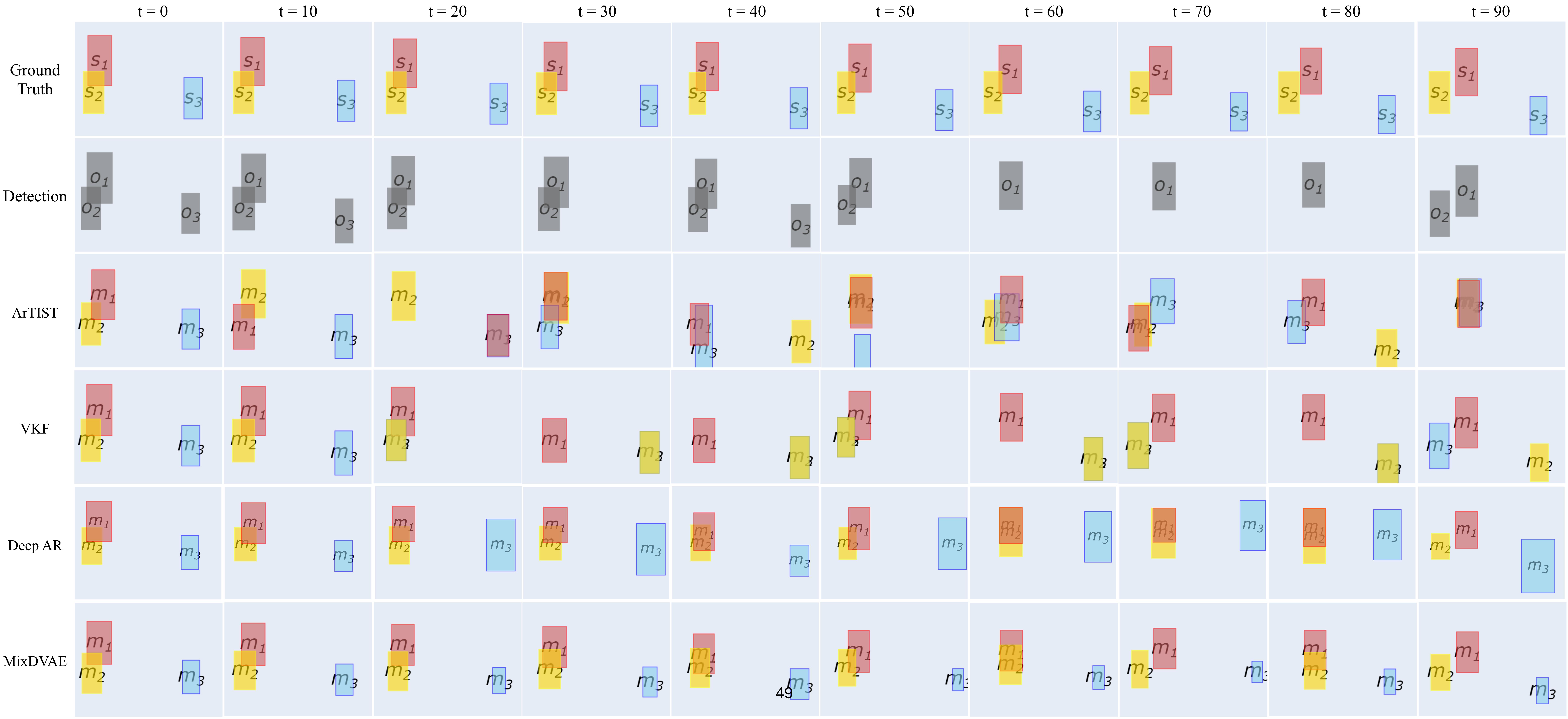
ArTIST (Saleh et al., 2021), VKF (Ban et al., 2020), Deep AR

Comparison with the SoTA models

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

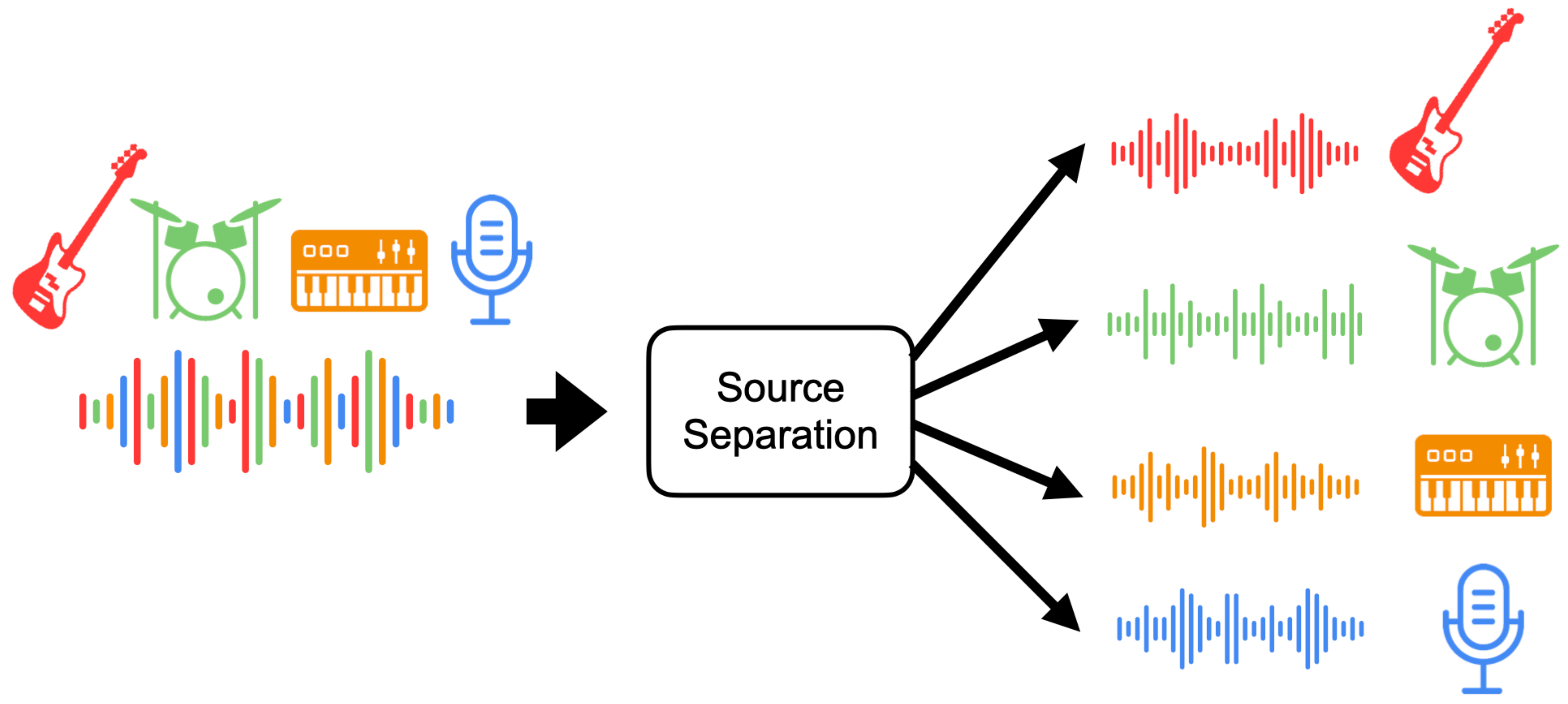
Dataset	Method	MOTA↑	MOTP↑	IDF1↑	#IDS↓	%IDS↓	MT↑	ML↓	#FP↓	%FP↓	#FN↓	%FN↓
Short	ArTIST	63.7	**84.1**	48.7	86371	28.0	**4684**	**0**	**9962**	**3.2**	**15525**	**5.0**
	VKF	56.0	82.7	77.3	5660	1.8	3742	761	64945	21.1	64945	21.1
	Deep AR	67.4	76.1	83.1	5248	1.7	3670	129	49595	16.0	49595	16.0
	MixDVAE	**79.1**	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

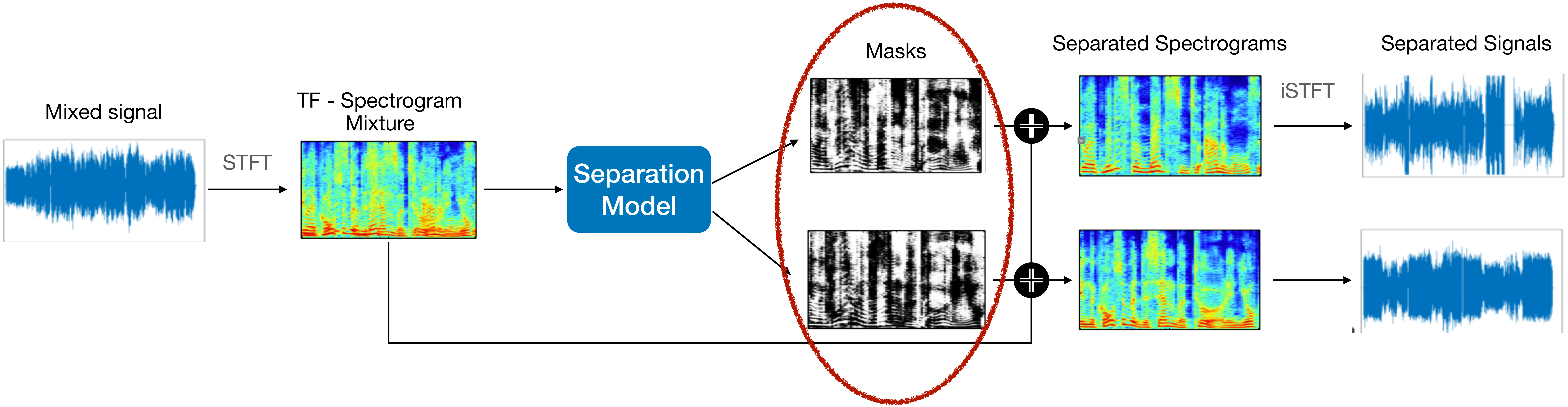


“Cocktail Party Effect” — Bregman 1990

Applications

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

SC-ASS: Time-Frequency Masking with probabilistic models

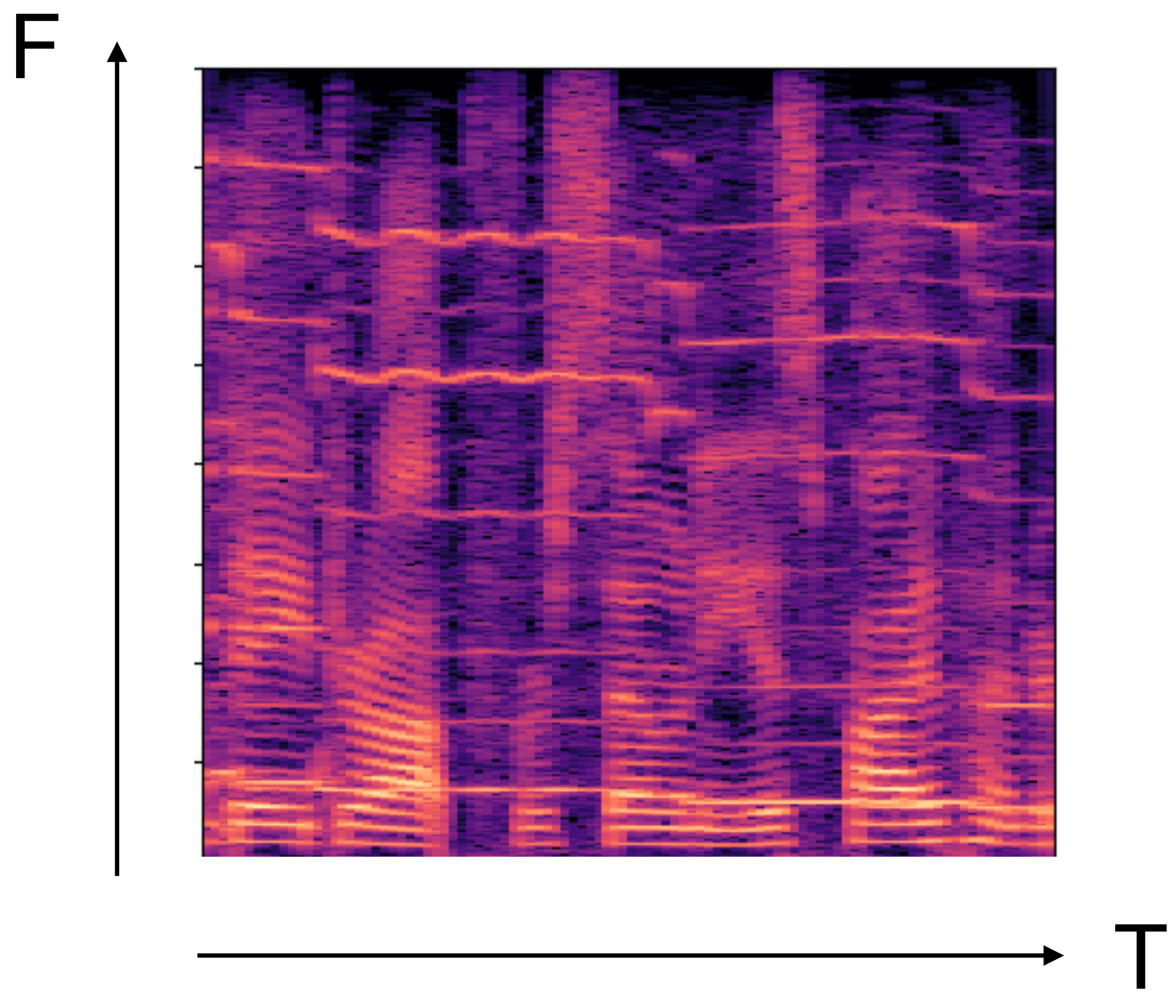


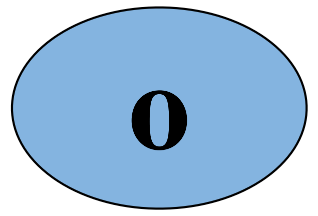
Key question: how to obtain the masks?

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

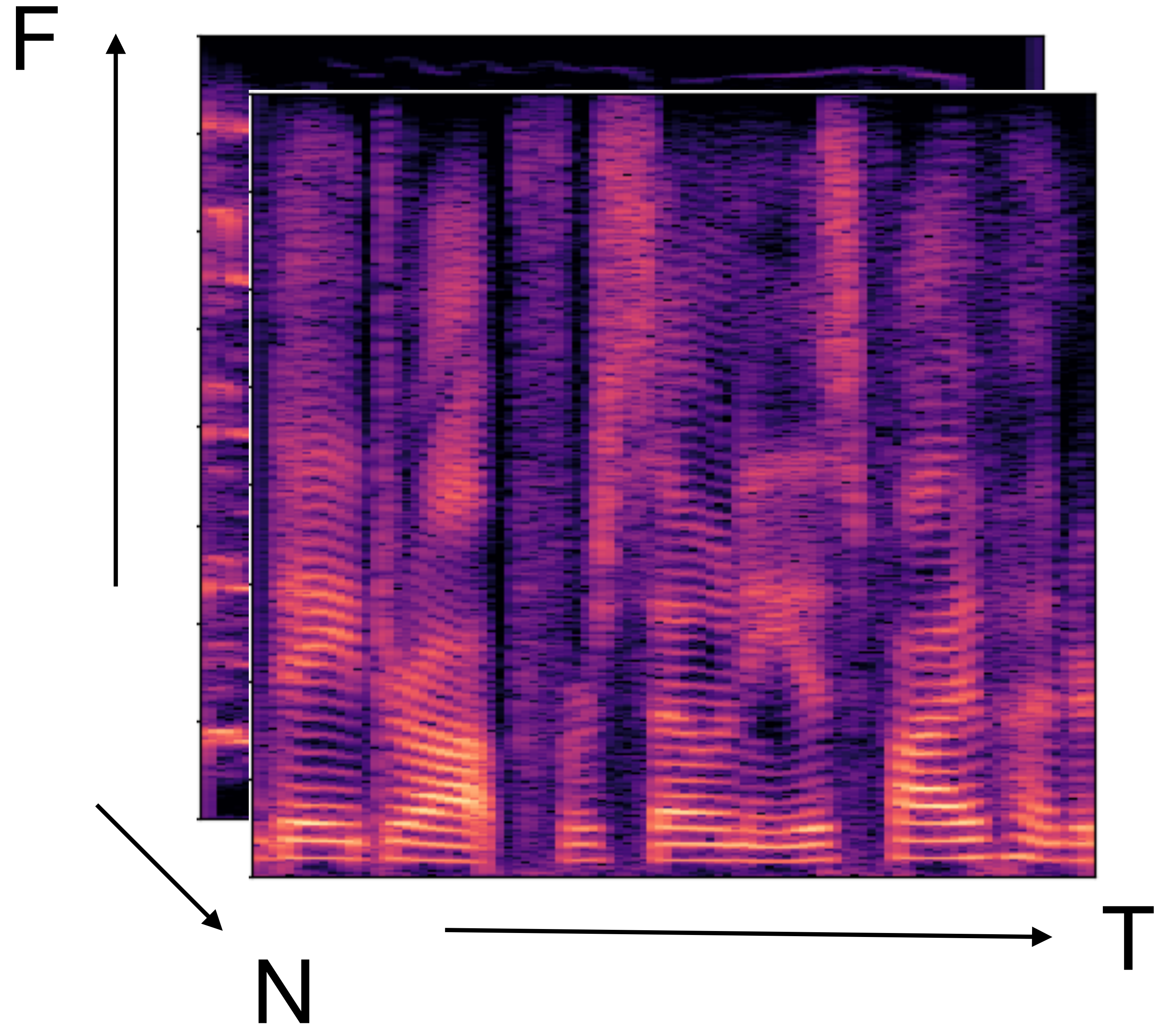


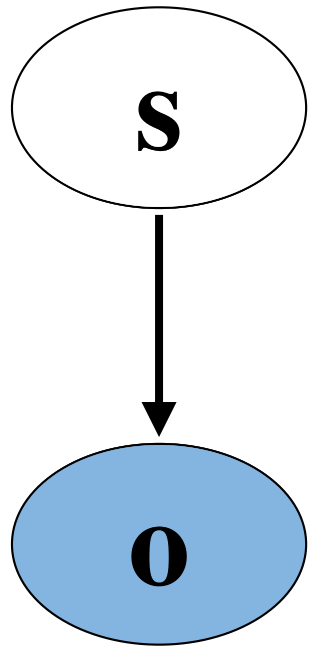


Define SC-ASS from a probabilistic perspective

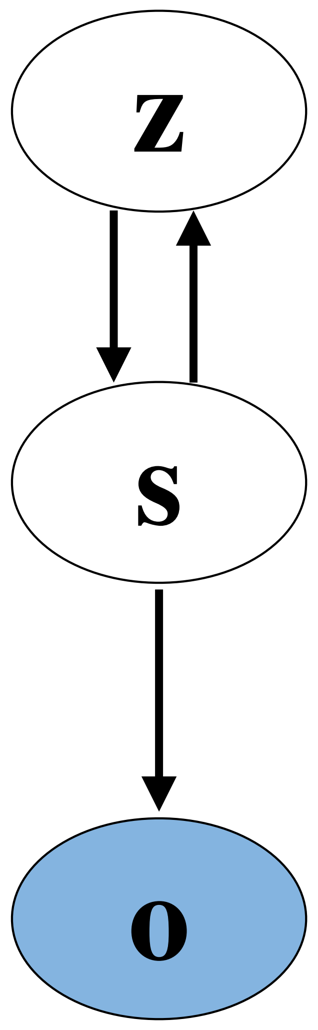
Definition of random variables

- $\mathbf{o} = \{o_{1:T,1:F}\} \in \mathbb{C}^{T \times F}$: 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





Define SC-ASS from a probabilistic perspective

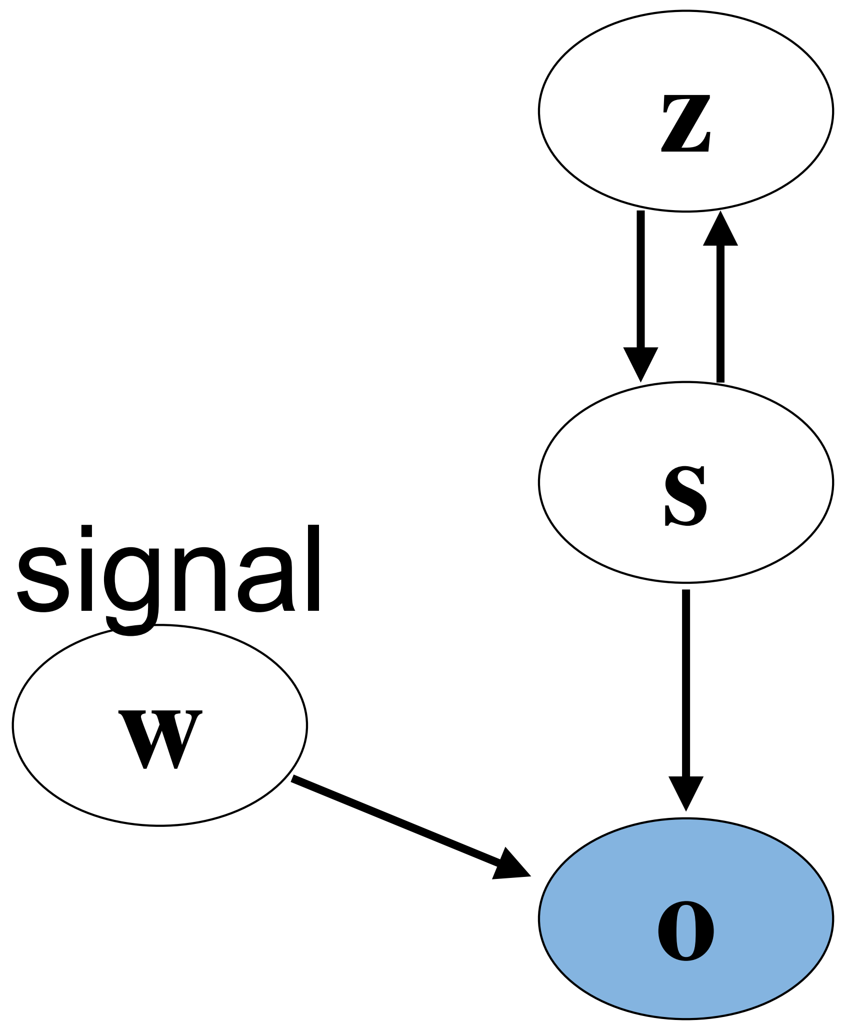


Definition of random variables

- $\mathbf{o} = \{o_{1:T,1:F}\} \in \mathbb{C}^{T\times F}$: 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

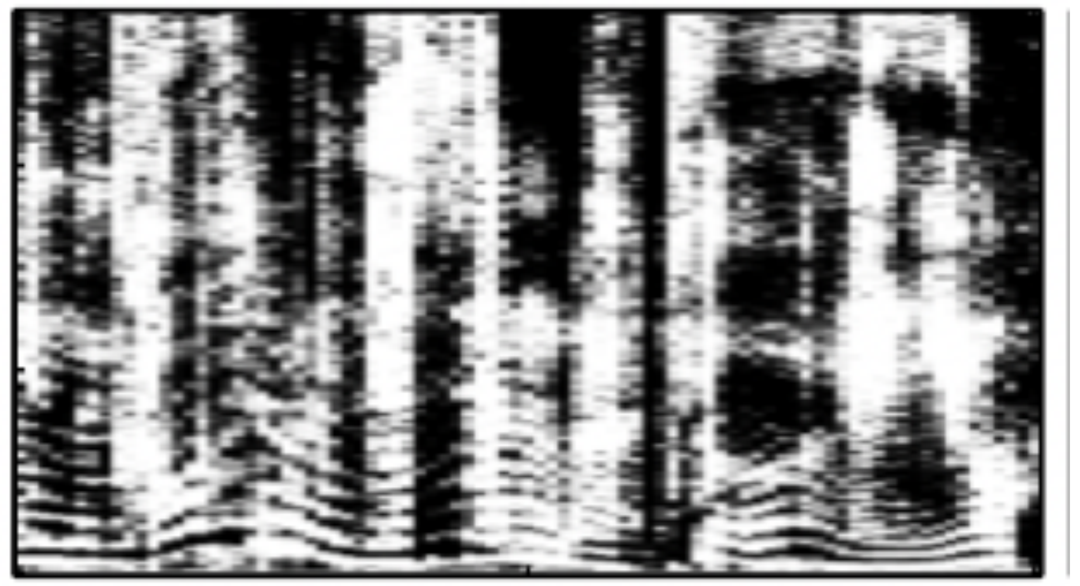
  $\mathbf{z} = \{\mathbf{z}_{1:N,1:T}\} \in \mathbb{R}^{N\times T\times L}$: latent sequences of DVAE models

Define SC-ASS from a probabilistic perspective

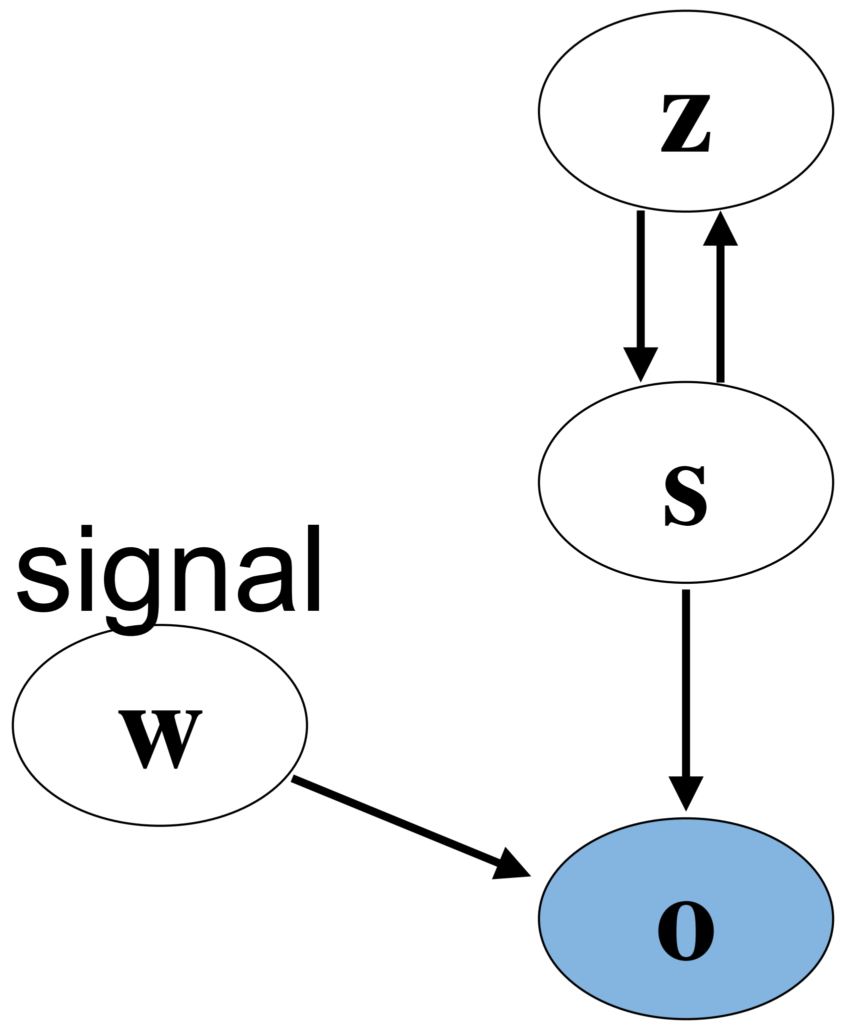


Definition of random variables

- $\mathbf{o} = \{o_{1:T,1:F}\} \in \mathbb{C}^{T\times F}$: 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



Define SC-ASS from a probabilistic perspective



Definition of random variables

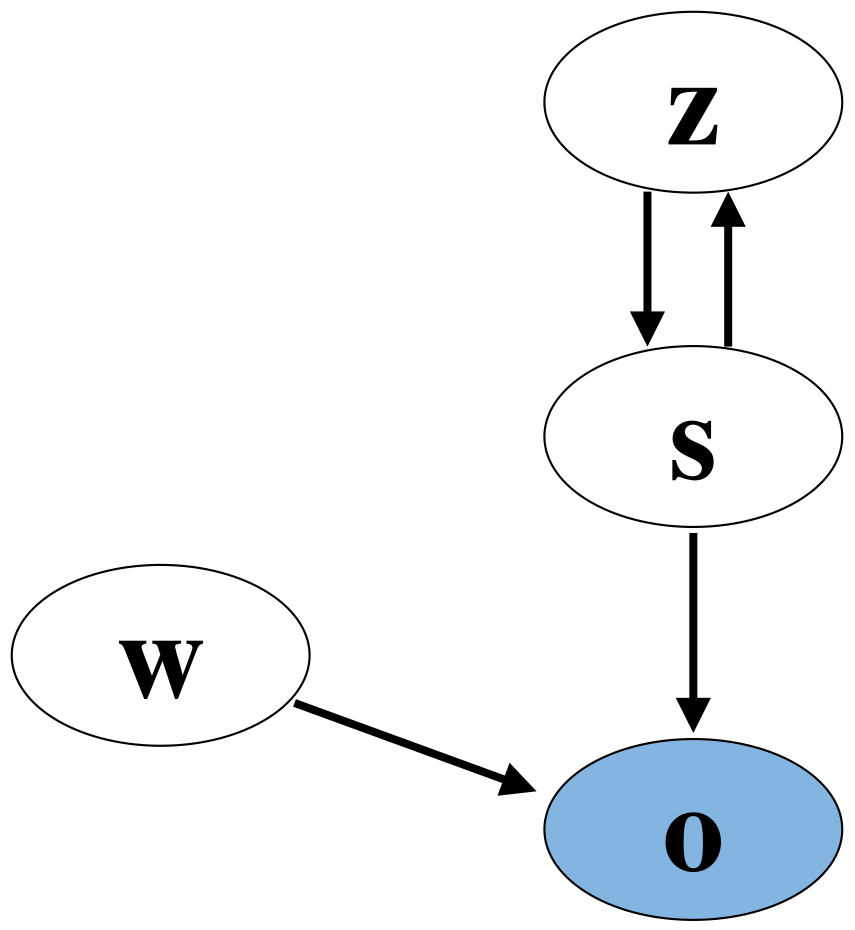
- $\mathbf{o} = \{o_{1:T,1:F}\} \in \mathbb{C}^{T\times F}$: STFT spectrogram of the observed mixture signal
- $\mathbf{s} = \{s_{1:N,1:T,1:F}\} \in \mathbb{C}^{N\times T\times F}$: STFT spectrograms of N sources
  $\mathbf{z} = \{\mathbf{z}_{1:N,1:T}\} \in \mathbb{R}^{N\times T\times L}$: latent sequences 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: $\mathbf{s}, \mathbf{z}, \mathbf{w}$

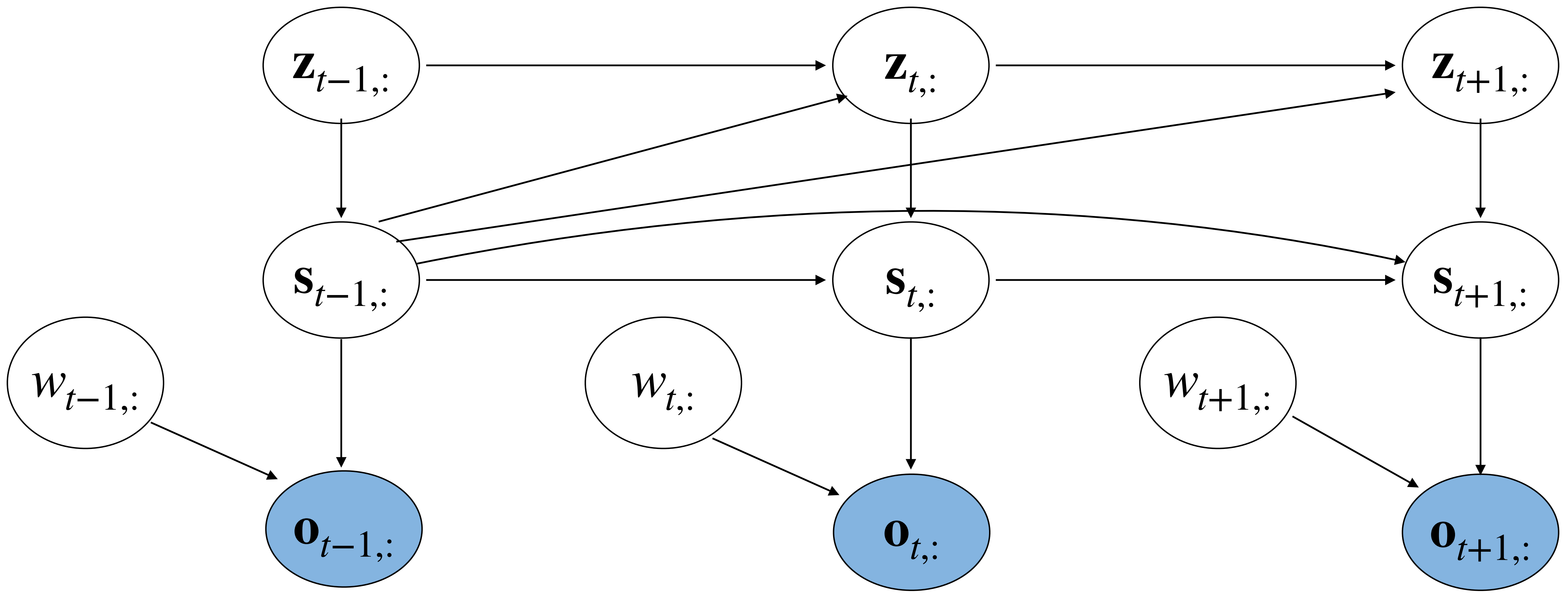
SC-ASS objective: estimate the posterior distribution $p(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

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.

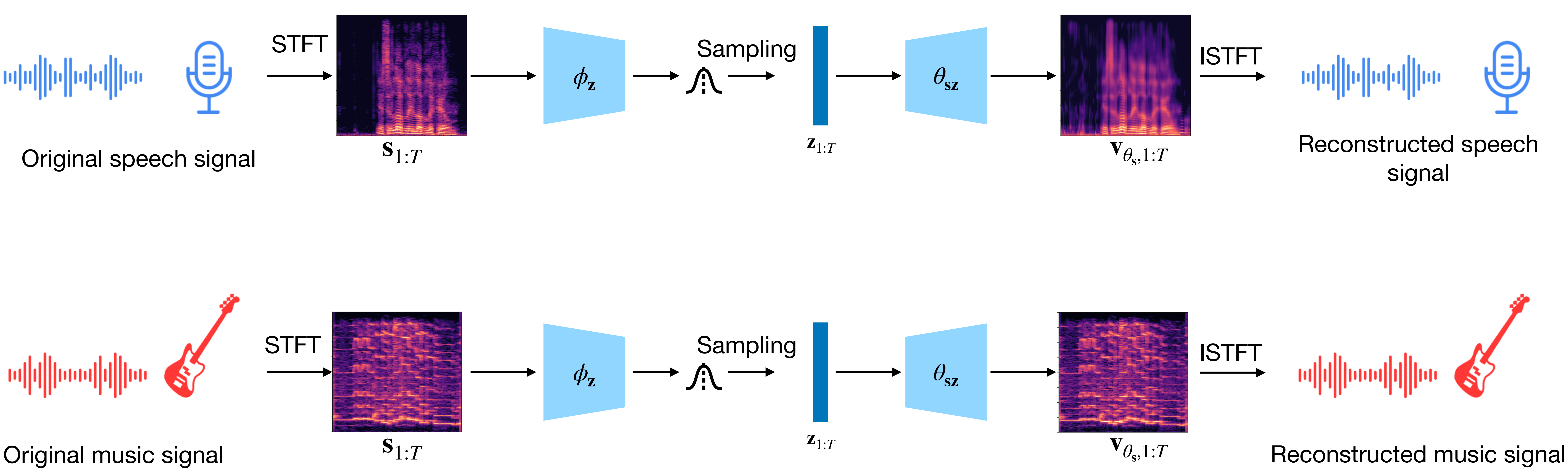
Intractable true posterior distribution $p_{\theta_\mathbf{szw}}(\mathbf{s}, \mathbf{z}, \mathbf{w} \mid \mathbf{o})$

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

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)

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

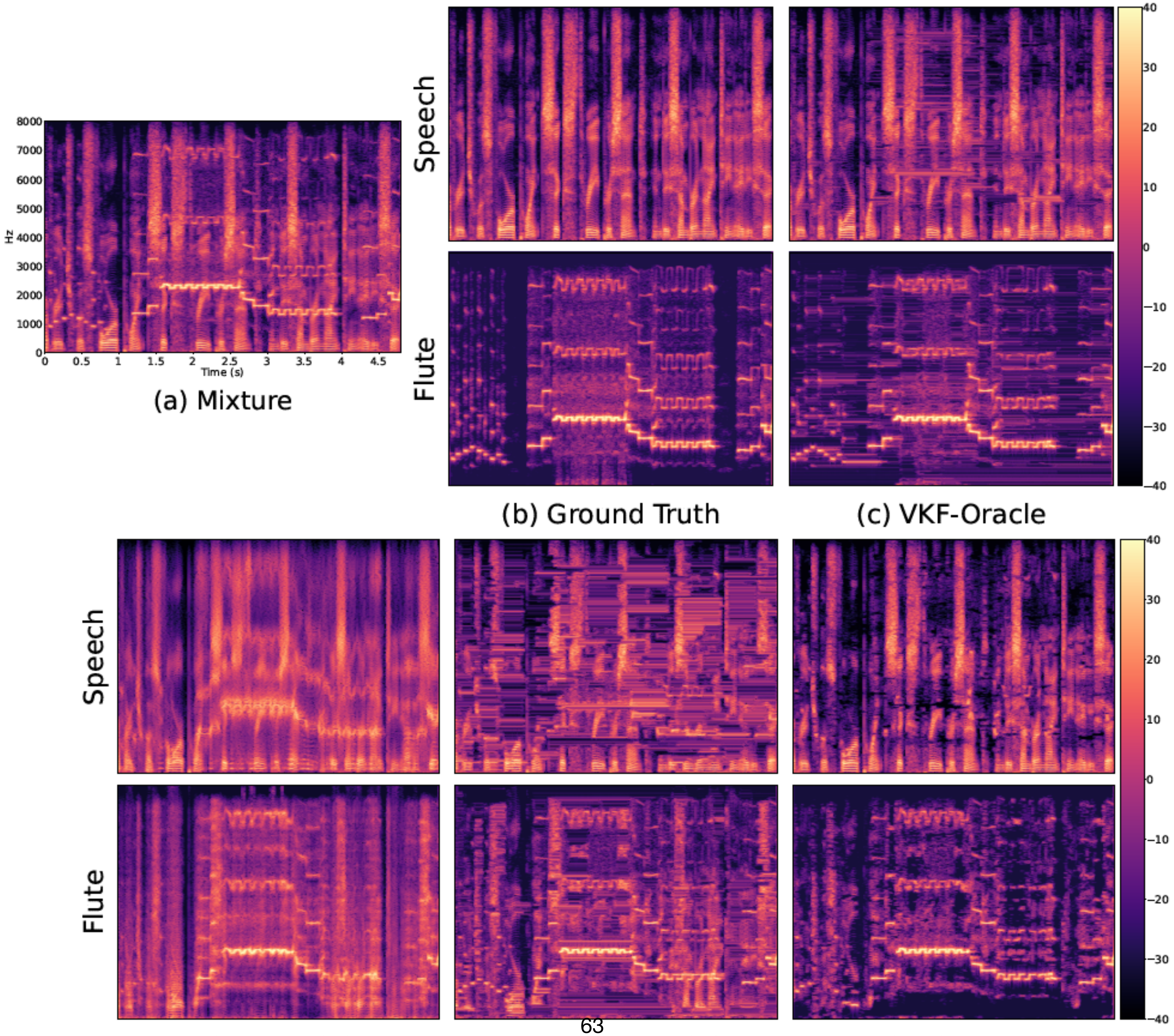
VKF, Deep AR, MixIT (Wisdom et al., 2020), Vanilla NMF (Févotte et al., 2018), temporal NMF (Virtanen, 2007)

Comparison with baseline models

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

Dataset	Method	Speech RMSE ↓	Speech SI-SDR ↑	Speech PESQ ↑	Chinese bamboo flute RMSE ↓	Chinese bamboo flute SI-SDR ↑	Chinese bamboo flute 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

Q & A