AA 598B Special Topics

Decision-Making & Control for Safe Interactive Autonomy

Instructor: Prof. Karen Leung

Autumn 2024

https://faculty.washington.edu/kymleung/aa598/

Announcements

- Guest lecture next Wednesday by Dr. Boris Ivanovic, Senior Research Scientist and Manager in NVIDIA Autonomous Vehicle Research Group
  - Submit talk review/reflection
- Start thinking about project proposals
  - Due Nov 1 Friday
- Homework #1 recommended due date this Friday
- https://lib.uw.edu/help/connect/tools/

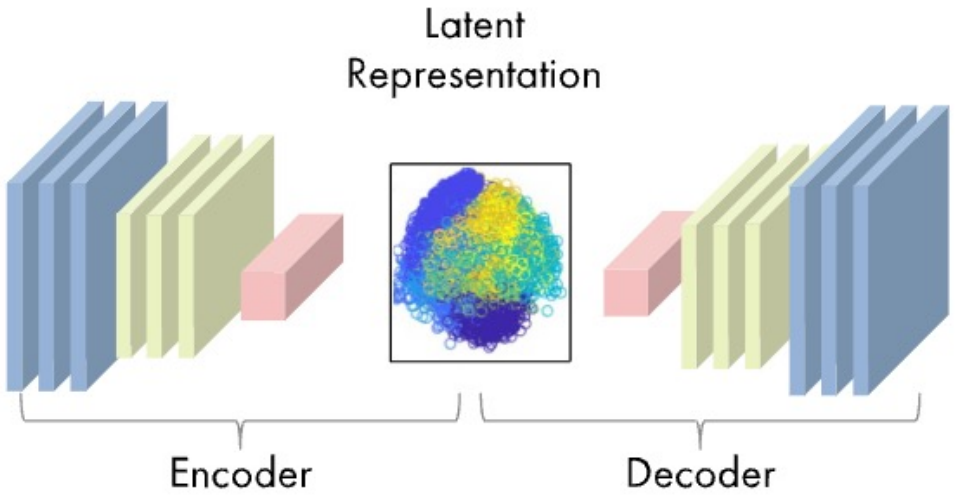
Last time

- Wrapped up ontological methods
  - Pros and Cons
- Discussed phenomenological methods (aka deep generative models)
  - Different types of generative models
  - Started talking about CVAEs

Today

- Wrap up CVAEs & latent space models
- Current landscape of prediction models
- Quickfire paper summary

Autoencoders



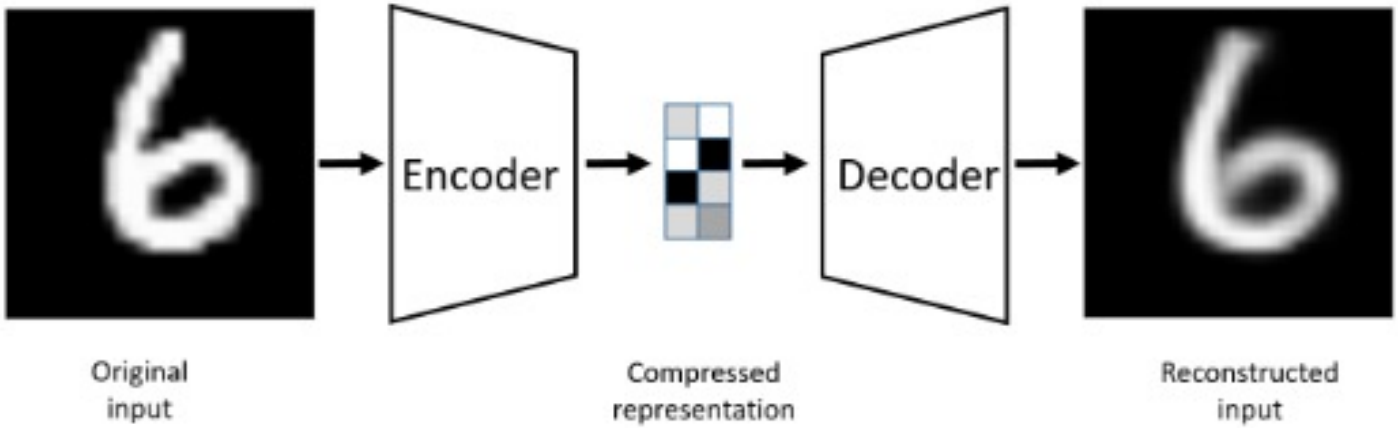
High dimensional data

Reconstruct input

Enforce a bottleneck via a low-dimensional layer

An approach to perform clustering, especially on high-dimensional data

Image compressions

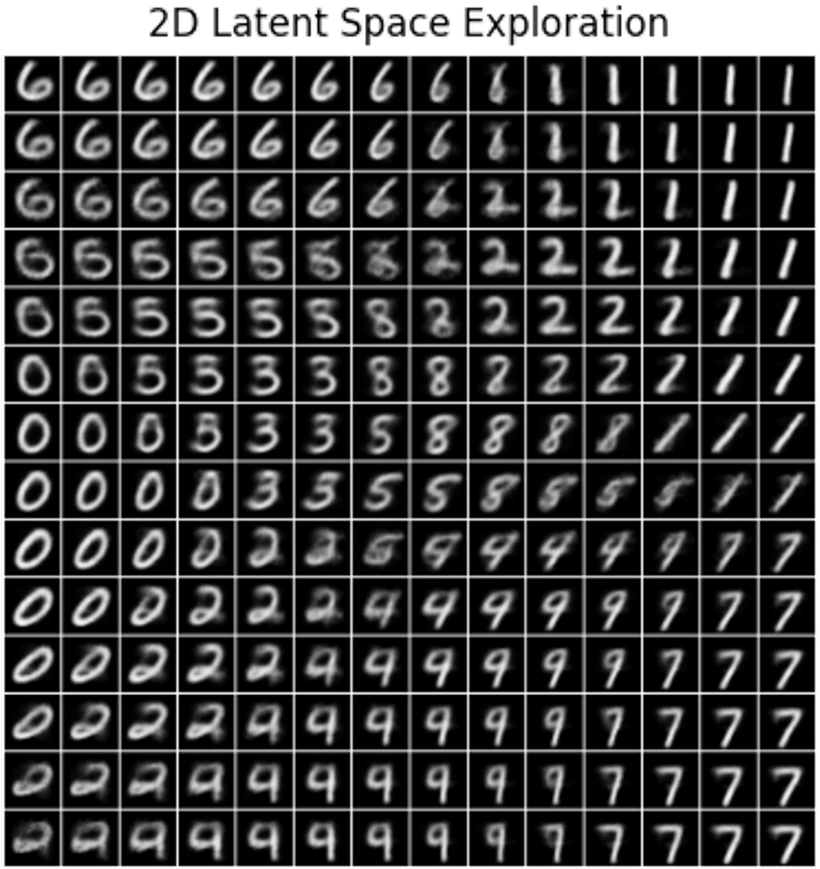


$x \longrightarrow q(z|x) \longrightarrow x$

$p(z)$

Murat et al 2023

MNIST dataset



https://arxiv.org/abs/1312.6114

https://kikaben.com/vae-2013/

Conditional Variational Autoencoders

want to learn → $p(y \mid x) = \int p(y \mid x, z) p(z \mid x) dz$

$p(y, z \mid x)$

z: latent variable

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(y_i \mid x_i) \Rightarrow \max_{\phi, \psi} \sum_{i=1}^{N} \int p_\phi(y_i \mid x_i, z) p_\psi(z \mid x_i) dz$$

intractable to compute

Evidence lower bound $\log p(y \mid x) \geq \mathbb{E}_{q(z \mid x, y)}[\log p(y \mid x, z)] - D_{\mathrm{KL}}(q(z \mid x, y) \mid p(z \mid x))$

ELBO

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

ϕ, ψ, φ \varpsi

Conditional Variational Autoencoders

output, cond. var

$$p(y \mid x) = \int p(y \mid x, z) p(z \mid x) dz$$ — marginalizing out z.

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(y_i \mid x_i) \Rightarrow \max_{\phi,\psi} \sum_{i=1}^{N} \int p_\phi(y_i \mid x_i, z) p_\psi(z \mid x_i) dz$$

z encodes meaning info of x

max likelihood

intractable! :(

consider $p_\psi(z|x)$, some z may be not meaningful ⇒ some z have very low probability. a lot of regions in $p(z|x)$ is zero.

(I want ψ to be similar.)

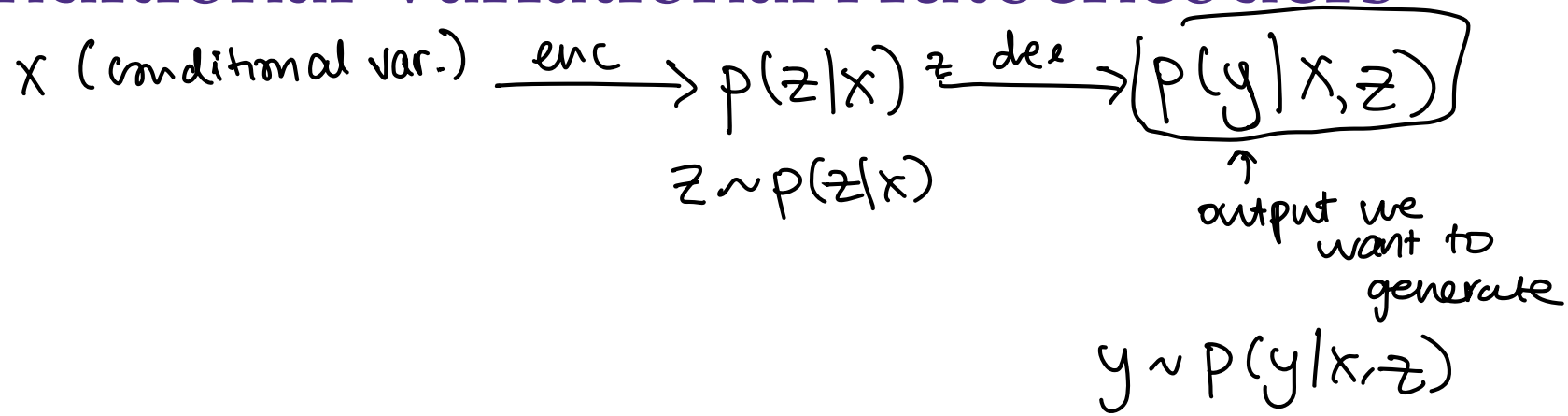
introduce $q_\varphi(z|x,y)$ → z that are likely given x, y.

Evidence lower bound $\log p(y \mid x) \geq \mathbb{E}_{q(z|x,y)}[\log p(y \mid x, z)] - D_{\mathrm{KL}}(q(z \mid x, y) \mid p(z \mid x))$

log likelihood term

encourage $q(z|x,y)$ to be similar to $\phi(z|x)$

Conditional Variational Autoencoders



$w = \mu + \sigma z \quad z \sim \mathcal{N}(0, I)$

Robot conditioned human trajectory predictor

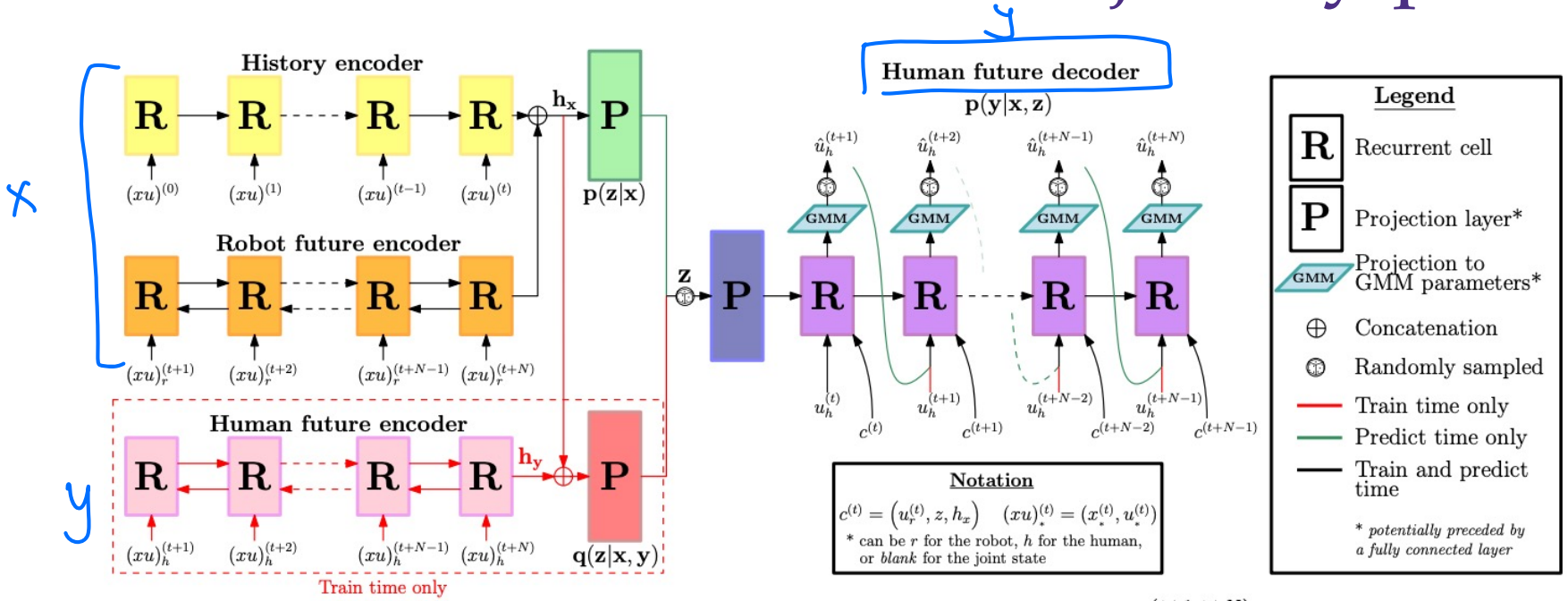
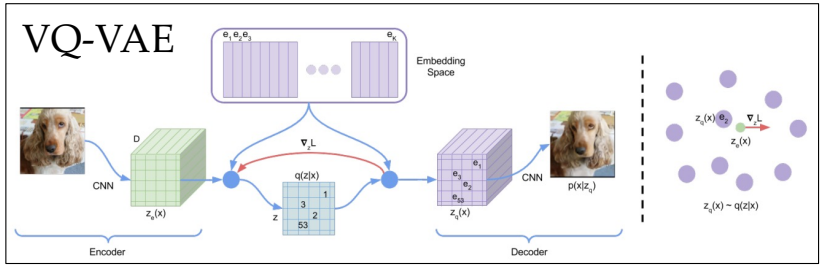


Fig. 2. CVAE architecture for sequence-to-sequence generative modeling of future human actions $\mathbf{y} = u_h^{(t+1:t+N)}$ conditioned on joint interaction history $(x^{(0:t)}, u^{(0:t)})$ and candidate robot future actions $u_r^{(t+1:t+N)}$ (together, $\mathbf{x}$). The random variable $\mathbf{z}$ is a latent mixture component index.

Schmerling et al 2018

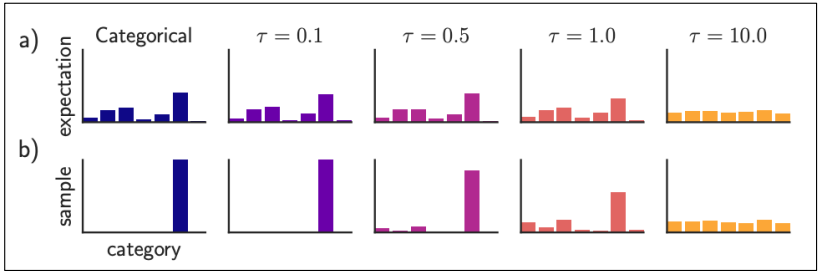
quantizing

Types of latent spaces



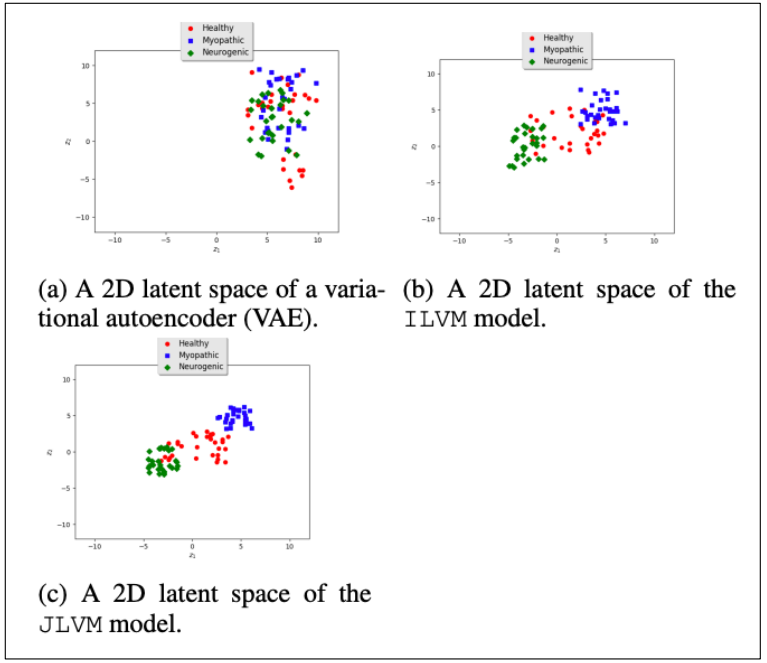
Van den Oord et al 2017

Categorical reparameterization



Jang et al 2017

Discovering interpretable representations



(a) A 2D latent space of a variational autoencoder (VAE).

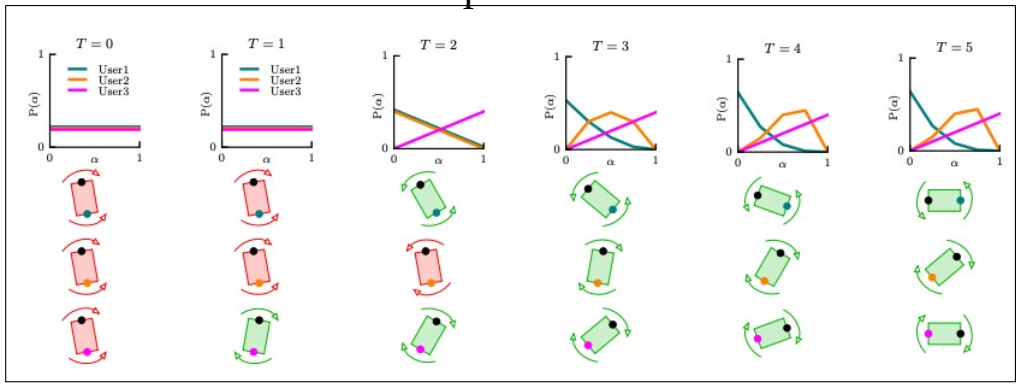
(b) A 2D latent space of the ILVM model.

(c) A 2D latent space of the JLVM model.

Adel et al 2018

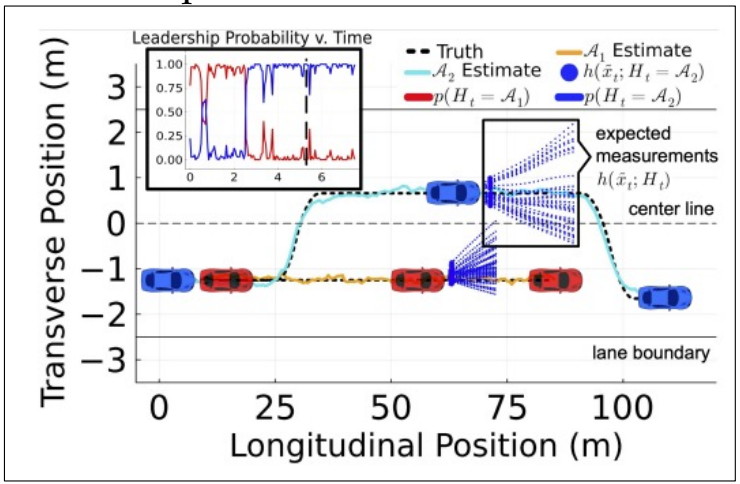
Types of latent spaces

Human-robot mutual adaptation



Nikolaidis et al 2015

Leadership inference



Khan & Fridovich-Keil 2024

Limitations of (C)VAEs

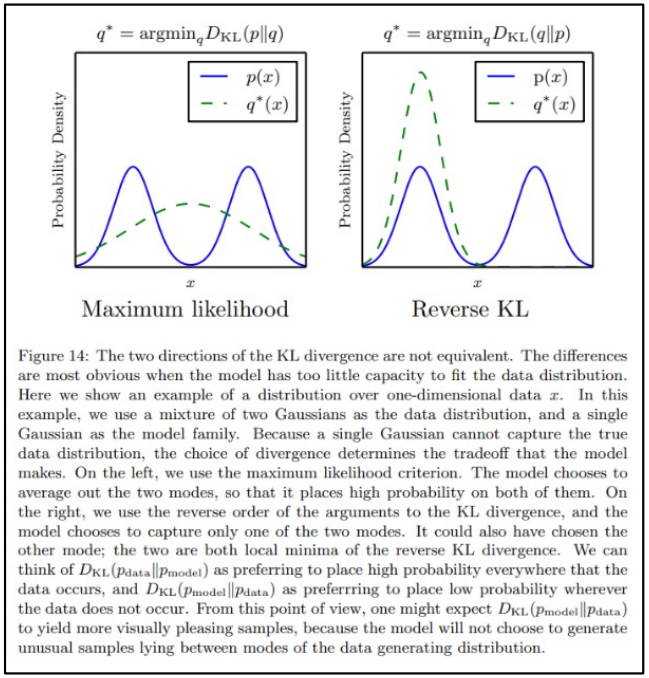
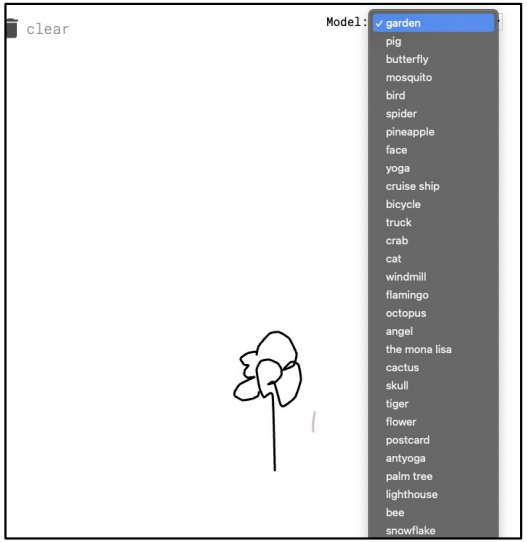


Figure 14: The two directions of the KL divergence are not equivalent. The differences are most obvious when the model has too little capacity to fit the data distribution. Here we show an example of a distribution over one-dimensional data x. In this example, we use a mixture of two Gaussians as the data distribution, and a single Gaussian as the model family. Because a single Gaussian cannot capture the true data distribution, the choice of divergence determines the tradeoff that the model makes. On the left, we use the maximum likelihood criterion. The model chooses to average out the two modes, so that it places high probability on both of them. On the right, we use the reverse order of the arguments to the KL divergence, and the model chooses to capture only one of the two modes. It could also have chosen the other mode; the two are both local minima of the reverse KL divergence. We can think of $D_{\text{KL}}(p_{\text{data}}\|p_{\text{model}})$ as preferring to place high probability everywhere that the data occurs, and $D_{\text{KL}}(p_{\text{model}}\|p_{\text{data}})$ as preferrring to place low probability wherever the data does not occur. From this point of view, one might expect $D_{\text{KL}}(p_{\text{model}}\|p_{\text{data}})$ to yield more visually pleasing samples, because the model will not choose to generate unusual samples lying between modes of the data generating distribution.

KL divergences can result in "blurry" reconstructions.

Posterior collapse leads to loss of meaningful latent space representations



Cannot add new conditioning variables after training

Other considerations for trajectory prediction

- Predicting discrete modes, e.g., homotopy classes
- Predicting high-level actions, e.g., goals
- Grouping agents
- Conditioning on agent class, environment, context
- Injecting rules, e.g., signal temporal logic, safety considerations

Behavior prediction models

Method	Road Enc.	Motion Enc.	Interactions	Decoder	Output	Trajectory Distribution
Jean [34]	–	LSTM	attention	LSTM	states	GMM
TNT [54]	polyline	polyline	maxpool, attention	MLP	states	Weighted set
LaneGCN [32]	GNN	1D conv	GNN	MLP	states	Weighted set
WIMP [28]	polyline	LSTM	GNN+attention	LSTM	states	GMM
VectorNet [20]	polyline	polyline	maxpool, attention	MLP	states	Single traj.
SceneTransformer [35]	polyline	attention	attention	attention	states	Weighted set
GOHOME [21]	GNN	1D conv + GRU	GNN	MLP	states	heatmap
MP3 [11]	raster	raster	conv	conv	cost function	Weighted samples
CoverNet [37]	raster	raster	conv	lookup	states	GMM w/ dynamic anchors
DESIRE [30]	raster	GRU	spatial pooling	GRU	states	Samples
RoadRules [27]	raster	raster	conv	LSTM	states	GMM
SocialLSTM [1]	–	LSTM	spatial pooling	LSTM	states	Samples
SocialGan [24]	–	LSTM	maxpool	LSTM	states	Samples
MFP [46]	raster	GRU	RNNs+attention	GRU	states	Samples
MANTRA [33]	raster	GRU	–	GRU	states	Samples
PRANK [2]	raster	raster	conv	lookup	states	Weighted set
IntentNet [10]	raster	raster	conv	conv	states	Single traj.
SpaGNN [9]	raster	raster	GNN	MLP	state	Single traj.
Multimodal [14]	raster	raster	conv	conv	states	Weighted set
PLOP [5]	raster	LSTM	conv	MLP	state poly	GMM
Precog [41]	raster	GRU	multi-agent sim.	GRU	motion	Samples
R2P2 [40]	raster	GRU	–	GRU	motion	Samples
HYU_ACE [36]	raster	LSTM	attn	LSTM	motion	Samples
Trajectron++[44]	raster	LSTM	RNNs+attention	GRU	controls	GMM
DKM [13]	raster	raster	conv	conv	controls	Weighted set
MultiPath [45]	raster	raster	conv	MLP	states	GMM w/ static anchors
MultiPath++	polyline	LSTM	RNNs+maxpool	MLP	control poly	GMM

Table 1: A survey of recent work in behavior prediction, categorized by choice of road encoding, motion history encoding, agent interaction encoding, trajectory decoding, intrinsic output representation, and distribution over future trajectories.

MultiPath++, Varadarajan et al 2021

Prediction metrics

- Average displacement error (ADE)over k most likely predictions
- minADE
- Final displacement error (FDE) over k most likely predictions
- minFDE
- Best-of-N
- Miss rate at 2 meters over k
- …

NuScenes by Motional

Leaderboard

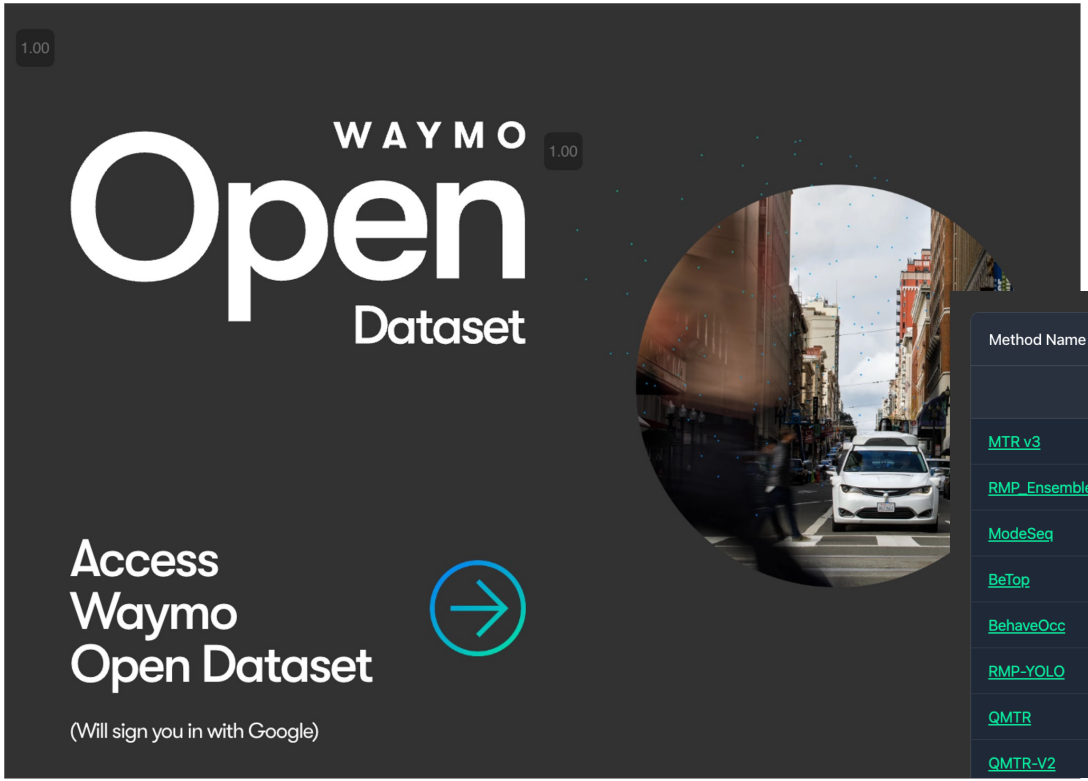
Search:

Export as JSON

Date	Name	MinADE_5	MinADE_10	MissRateTopK_2_5	MissRateTopK_2_10	MinFDE_1	OffRoadRate
2022-09-06	sub_test	1.430	1.037	65.70%	43.56%	8.703	0.036
2022-09-06	Jack	1.211	0.894	56.65%	33.38%	7.502	0.019
2022-09-02	HGO	1.701	1.701	67.37%	67.37%	9.021	0.064
2022-08-22	b	3.435	3.435	94.06%	94.06%	8.740	0.185
2022-08-01	Map_encoding_CNI	3.392	3.392	93.68%	93.68%	8.612	0.185
2022-07-25	MR_8_d	2.292	1.573	66.94%	54.73%	9.072	0.084
2022-07-24	MR	2.327	1.548	67.61%	56.28%	9.015	0.063
2022-07-20	dongdongteam	2.522	1.468	69.87%	55.81%	8.748	0.057

NuScenes prediction task

Waymo Open Dataset



Method Name	ition time	Soft mAP v2	mAP v2	minADE	minFDE	Miss rate
		Show results with				
MTR v3		0.4967	0.4859	0.5554	1.1062	0.1098
RMP_Ensemble		0.4737	0.4531	0.5564	1.1188	0.1084
ModeSeq		0.4737	0.4665	0.5680	1.1766	0.1204
BeTop		0.4698	0.4587	0.5716	1.1668	0.1183
BehaveOcc		0.4678	0.4566	0.5723	1.1668	0.1176
RMP-YOLO		0.4673	0.4523	0.5737	1.1697	0.1160
QMTR		0.4649	0.4445	0.5702	1.1627	0.1177
QMTR-V2		0.4646	0.4441	0.5700	1.1621	0.1174

Waymo Open Challenge – Motion Prediction

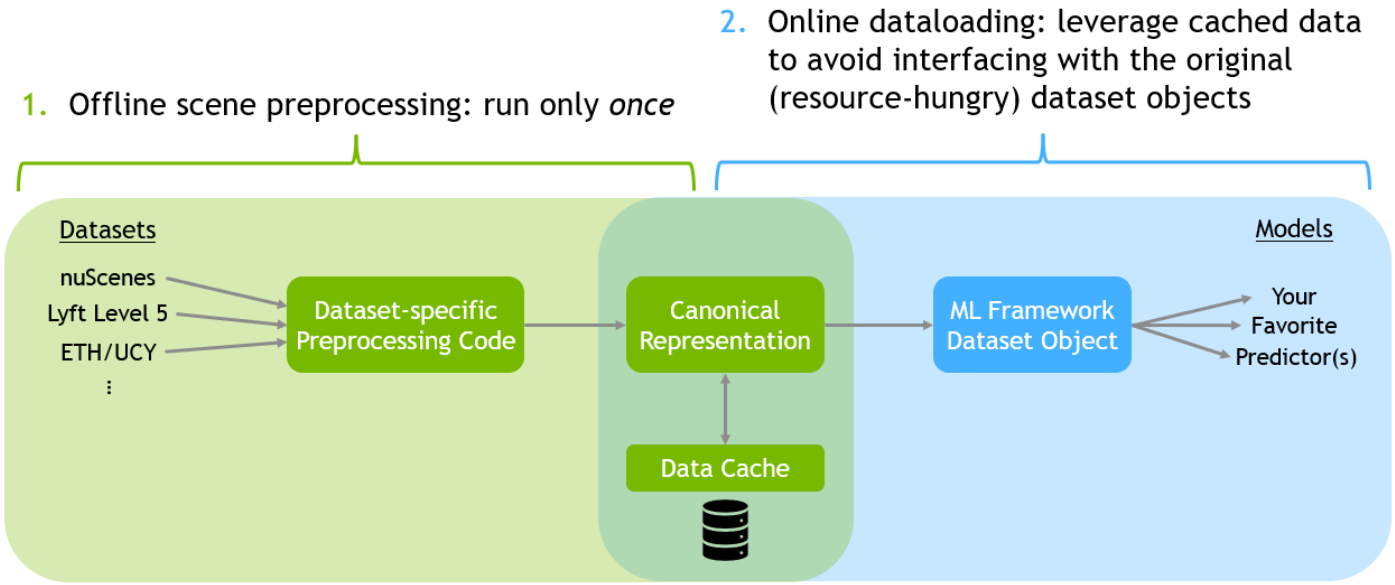
Stanford Drone Dataset

https://cvgl.stanford.edu/projects/uav_data/

Other datasets

- INTERACTION Dataset: INTERnational, Adversarial, and Cooperative motion Dataset
- ETH-UCY human navigation dataset
- THOR dataset: Human motion trajectories in indoor environments
- German dataset (highD, round, intD…)

TrajData

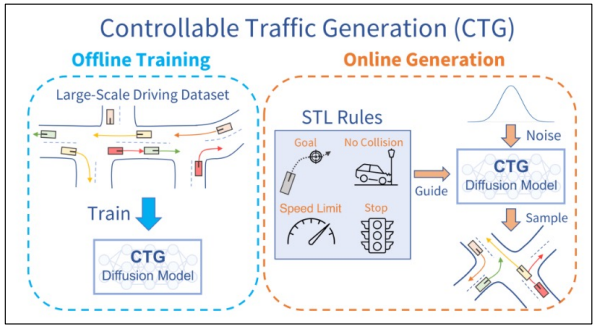


https://github.com/NVlabs/trajdata

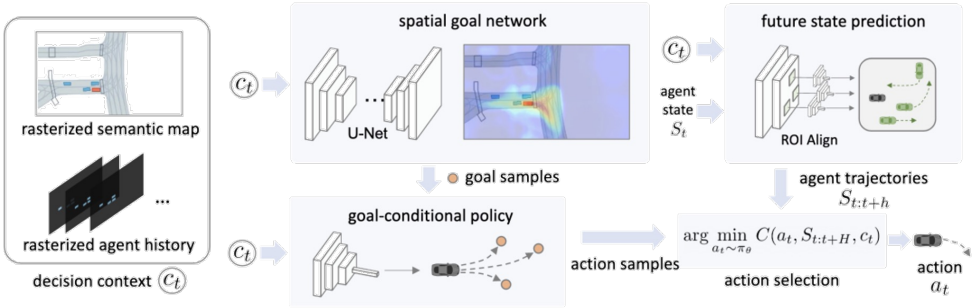
Other types of generative modeling applications

Controllable and compositional generation

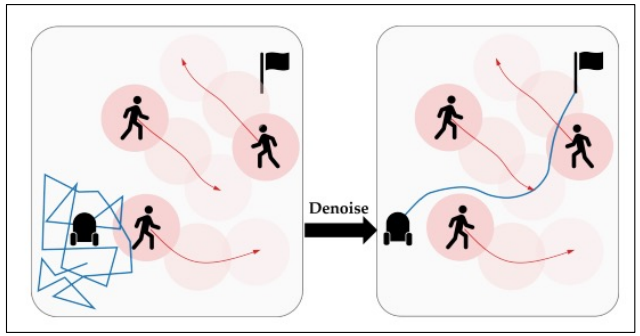
https://github.com/UW-CTRL/stljax



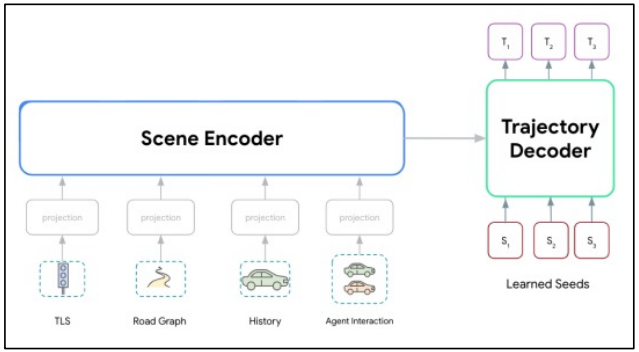
Guided Conditional Diffusion for Controllable Traffic Simulation



BITS: Bi-level Imitation for Traffic Simulation



CoBL-Diffusion: Diffusion-Based Conditional Robot Planning in Dynamic Environments Using Control Barrier and Lyapunov Functions



Wayformer: Motion Forecasting via Simple & Efficient Attention Networks

Quick fire paper summaries

- *Skim* a paper on human behavior prediction
- *Throw* together **1 slide** summarizing the paper
  - Representative image?
  - What is the *secret sauce*?
  - A single sentence describing the work
- For the class to share ☺

https://docs.google.com/presentation/d/15GOQXbULv5Q4kcYJu4ZiiW2ocMz6s505wL50RkfBXJU/edit?usp=sharing

Quick fire paper summaries

[1905.06113] Human Motion Trajectory Prediction: A Survey ←LOOK AT PAPERS CITED IN HERE
[2306.15136] What Truly Matters in Trajectory Prediction for Autonomous Driving?
[2111.14973v3] MultiPath++: Efficient Information Fusion and Trajectory Aggregation for Behavior Prediction
[2309.17209] Robots That Can See: Leveraging Human Pose for Trajectory Prediction
[2003.07847] PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling
[2303.10895] Leapfrog Diffusion Model for Stochastic Trajectory Prediction
[2305.17600] NashFormer: Leveraging Local Nash Equilibria for Semantically Diverse Trajectory Prediction