Composable Text Controls in Latent Space with ODEs

Guangyi Liu

The Chinese University of Hong Kong, Shenzhen
&
Mohamed Bin Zayed University of Artificial Intelligence

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

UC San Diego

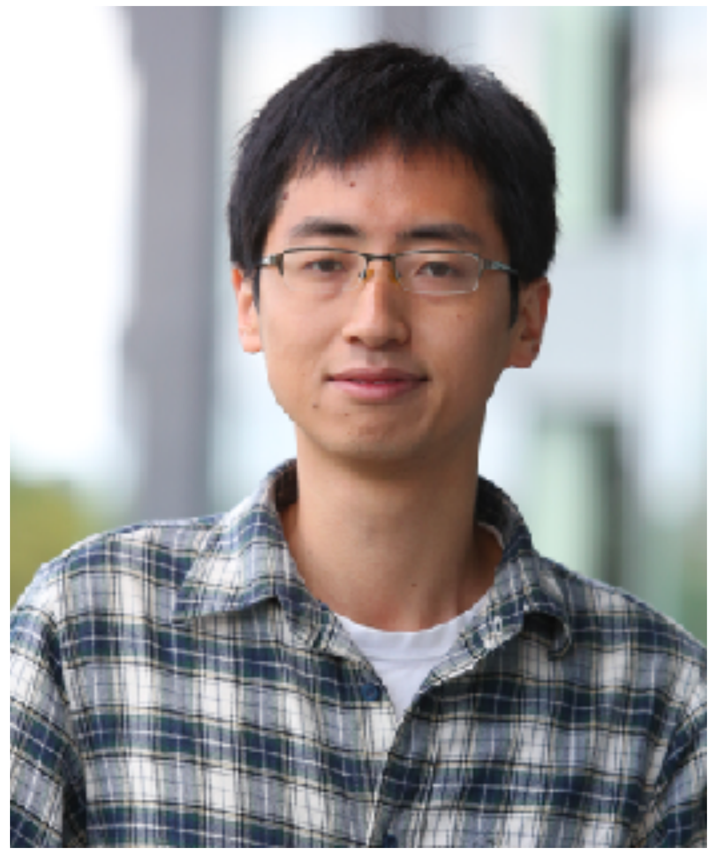
Carnegie Mellon University

Composable Text Controls in Latent Space with ODEs

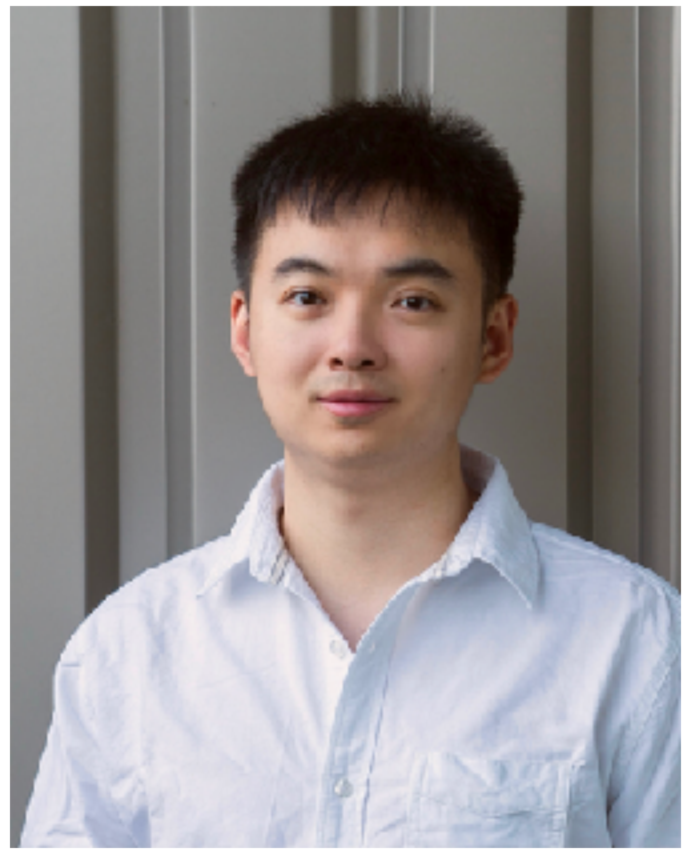
Guangyi Liu

Zeyu Feng

Yuan Gao



Zichao Yang



Zhiting Hu

Outline

- **Problem Statement**
- Background
- Method
  - Composable Latent-Space EBMs
  - Efficient Sampling with ODEs
  - Adapting Pretrained LMs for Latent Space
  - Implementation details
- Experiments
- Summary

Problem I

- Text Editing (e.g., Text Style Transfer)
  - Goal: **edit the attribute** of a given text and keep the **content preserved**.

Sentiment

Negative	I can't believe how ***awful*** this doctor is!	I can't believe how ***great*** this doctor is!	Positive

Tense

Present	I ***can't*** believe how awful this doctor ***is***!	I ***didn't*** believe how awful this doctor ***was***!	Past

- Plenty of works that can achieve very good performance on this specific task.
  - Adopt content loss, attribute loss and so on.

Problem I

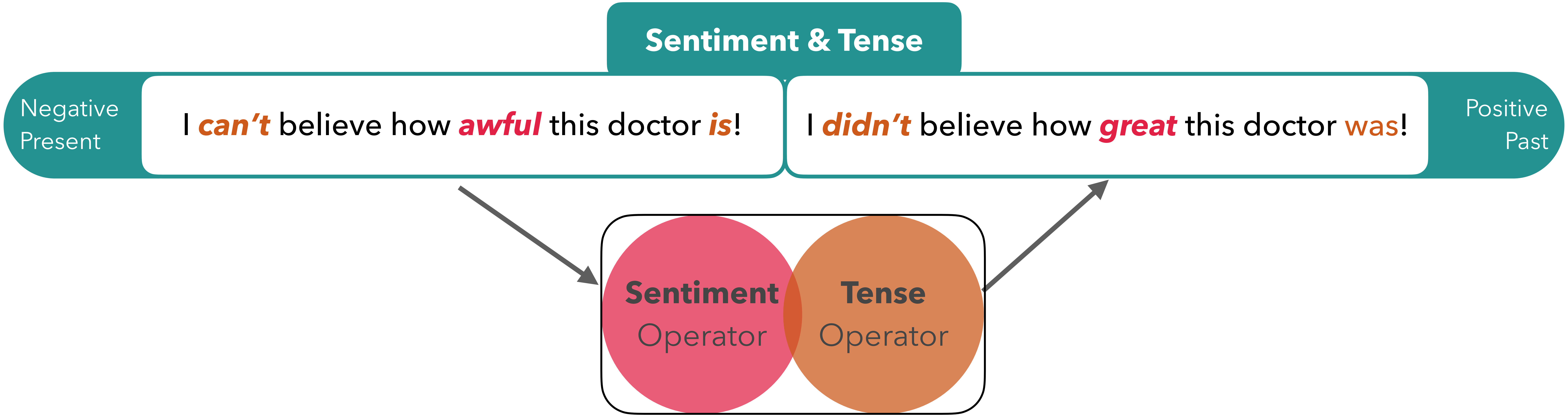
- Text Editing with Compositional Attributes
  - Example: Compose sentiment and tense:

Sentiment & Tense

Negative Present	I ***can't*** believe how ***awful*** this doctor ***is***!	I ***didn't*** believe how ***great*** this doctor was!	Positive Past

Problem I

- Text Editing with Compositional Attributes
  - Example: Compose sentiment and tense:



- Ideally solution: for each attribute, we have the corresponding **operator**, and these operators can be freely combined.

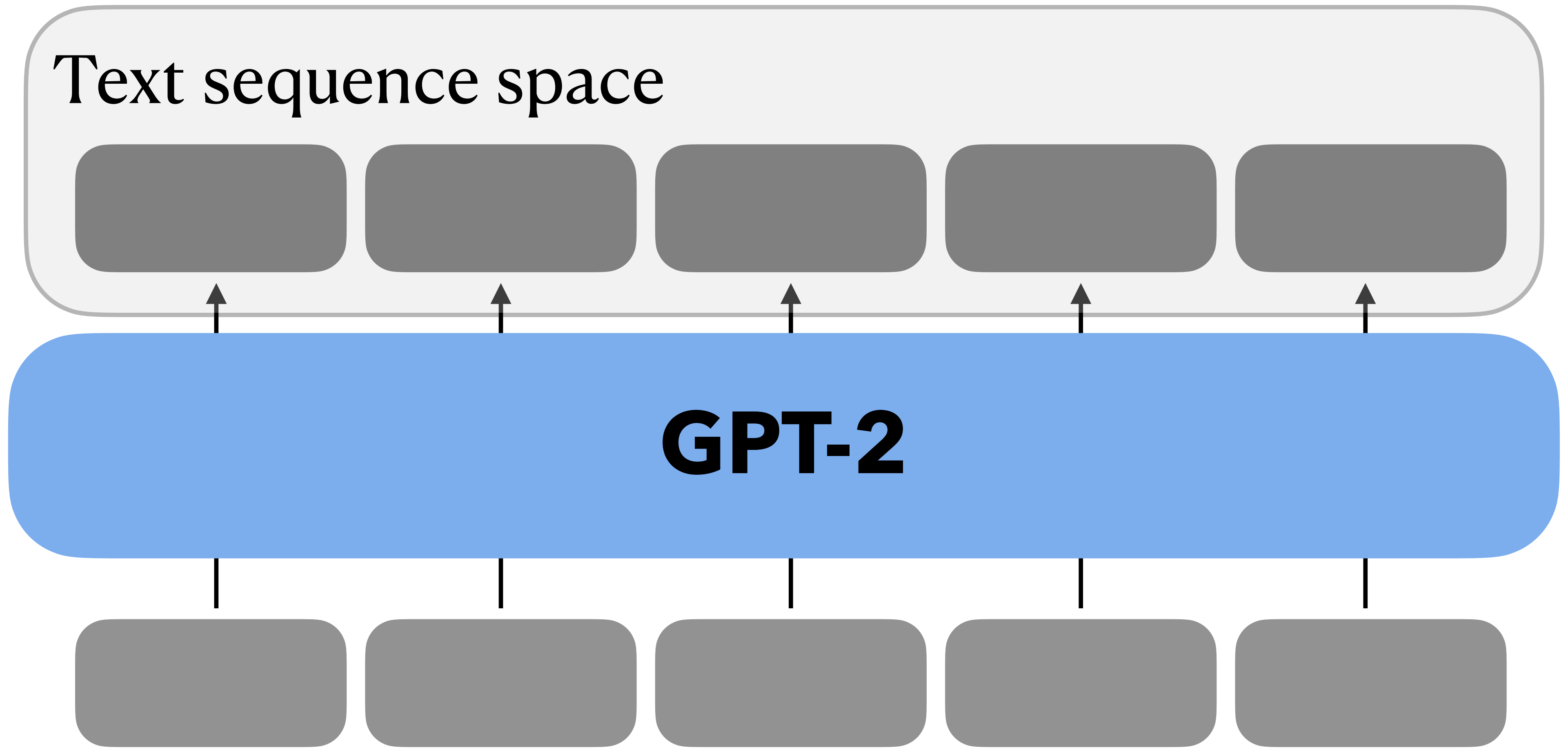
Problem II

- Conditional **Generation** with **Compositional Attributes**
  - Goal: generate **fluent** and **diverse** texts with **desired attributes**.

Problem II

- Conditional **Generation** with **Compositional Attributes**
  - Goal: generate **fluent** and **diverse** texts with **desired attributes**.
  - Some prior works (PLM-based, like **PPLM**[1] and **FUDGE**[2]), can guarantee the fluency.
  - **Diversity** and **accuracy** are still a problem.
  - Operate in the complex **text sequence space ->** **inefficient** generation

Lack of diversity:
1. **great** location.
2. **great**.
3. **great** place for lunch or a date.
4. **great** place!
5. **great** food.



[1]Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." *ICLR 2020*

[2]Yang, Kevin, and Dan Klein. "FUDGE: Controlled text generation with future discriminators." *NAACL 2021*

Solutions

- What we want:
  1. Good **Fluency**
  2. Good **Diversity**
  3. **Efficient** Generation
  4. **Compositionality**

- Possible solutions:
  1. PLMs, like **GPT-2**
  2. Strong Generative Models, like **VAEs**, GANs, DPM
  3. Operate in Low-Dimensional Latent Space
  4. Energy-Based Models are flexible to compose

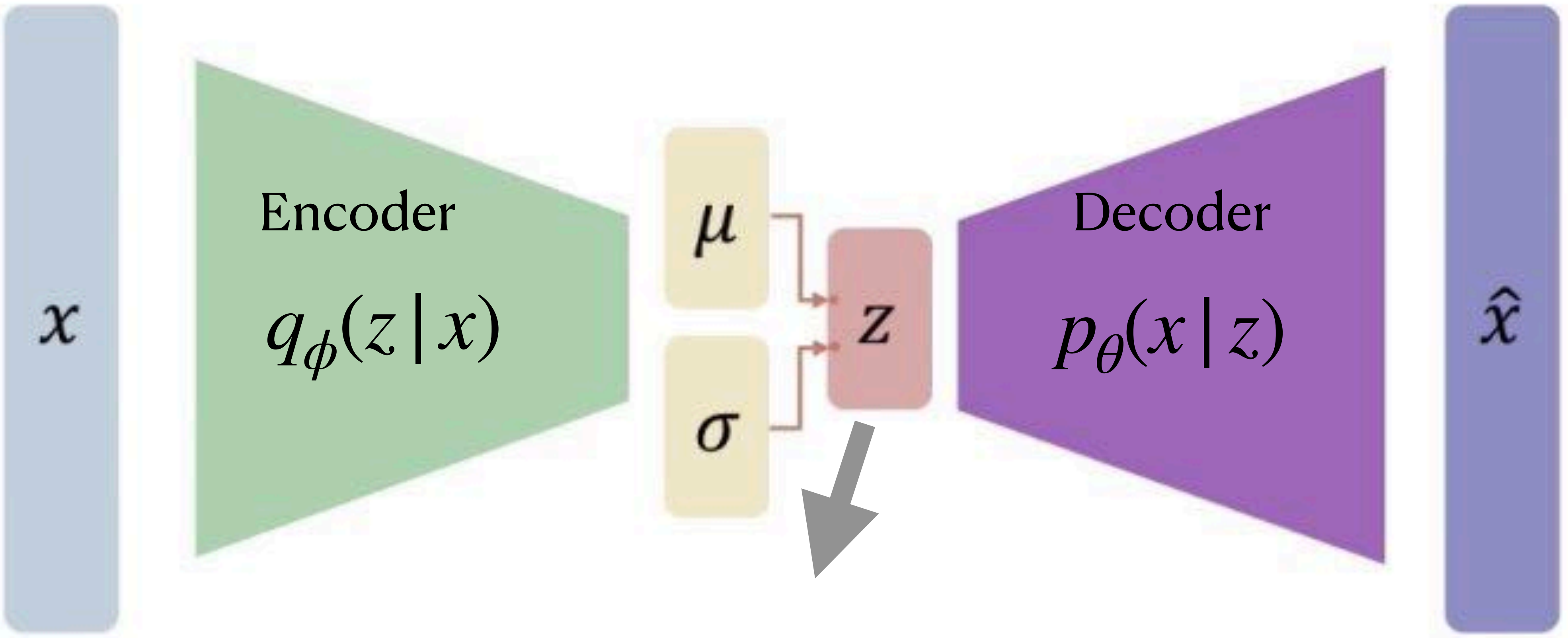
Outline

- Problem Statement
- **Background**
- Method
  - Composable Latent-Space EBMs
  - Efficient Sampling with ODEs
  - Adapting Pretrained LMs for Latent Space
  - Implementation details
- Experiments
- Summary

Background

Variational Auto-Encoders

x=“this is an example.”



$\hat{x}$=“this is an example.”

$$z \sim \mathcal{N}(\mu, \sigma) = q_\phi(z|x)$$

$$z \sim \mathcal{N}(0,I) = p_{\text{prior}}(z)$$

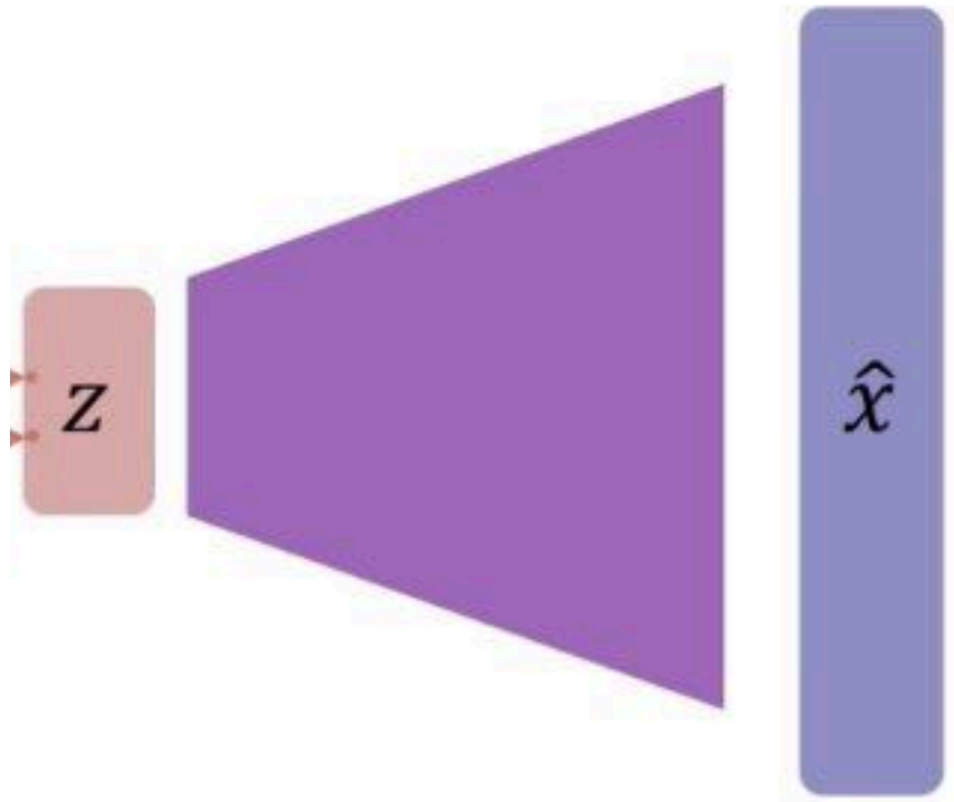


Figure from: https://towardsdatascience.com/reparameterization-trick-126062cfd3c3

Background

Energy-Based Generative Models

Given an arbitrary energy function $E(x) : \mathbb{R}^d \to \mathbb{R}$, energy-based models (EBMs) define a distribution:

$$p(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}/Z,$$

where $Z = \sum_{\boldsymbol{x} \in \mathcal{X}} e^{-E(\boldsymbol{x})}$ is the normalization term.

EBMs are flexible to incorporate any functions or constraints into the energy function $E(x)$.

[3] Song et al. "Score-based generative modeling through stochastic differential equations". *ICLR 2021.*

Background

Sampling from EBMs

- Langevin Dynamics is gradient-based MCMC approach
  - Sensitive to hyperparameters and unrobust in practice.

$$x_0 \sim p_0(x), \quad x_{t+1} = x_t - \frac{\eta}{2}\nabla_x E_\theta(x_t) + \epsilon_t, \quad \epsilon_t \sim N(0, \eta I)$$

- Stochastic Differential Equations [3] (SDEs):

$$\mathrm{d}\boldsymbol{x} = -\frac{1}{2}\beta(t)[\boldsymbol{x} + 2\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\bar{\boldsymbol{w}},$$

- Ordinary Differential Equations (ODEs):

$$\mathrm{d}\boldsymbol{x} = -\frac{1}{2}\beta(t)[\boldsymbol{x} + \nabla_x \log p_t(\boldsymbol{x})]\mathrm{d}t.$$

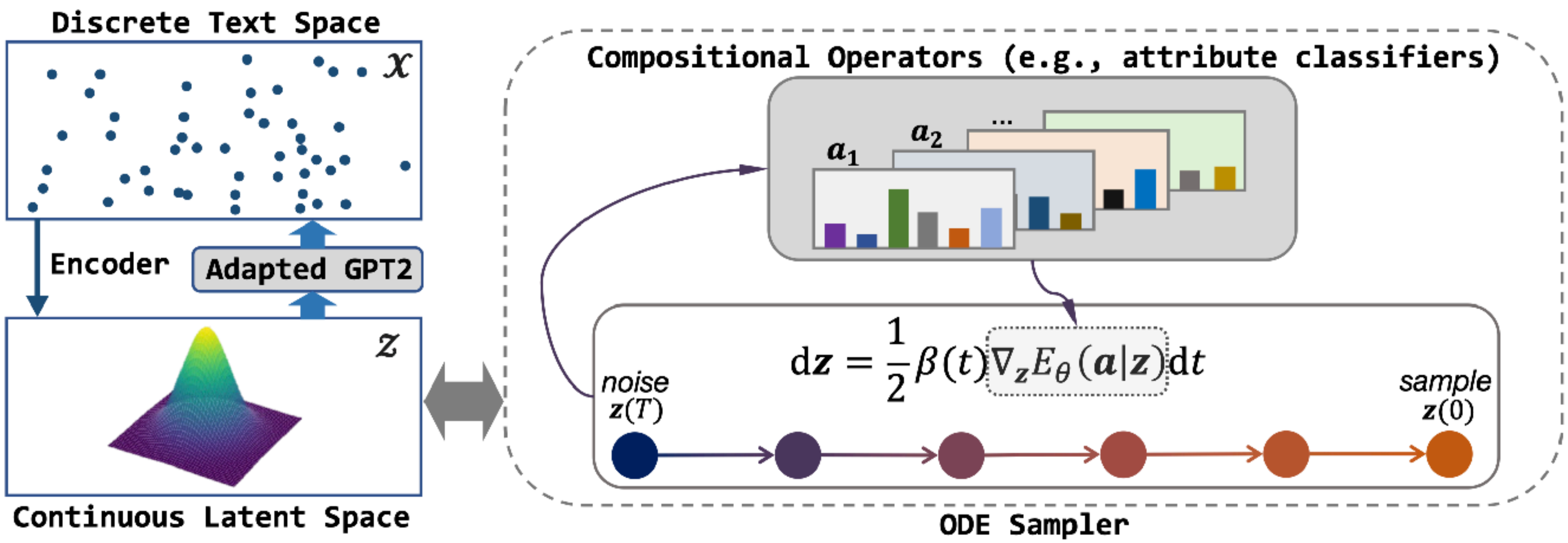
[3] Song et al. "Score-based generative modeling through stochastic differential equations". *ICLR 2021*.

Outline

- Problem Statement
- Background
- **Method**
  - Composable Latent-Space EBMs
  - Efficient Sampling with ODEs
  - Adapting Pretrained LMs for Latent Space
  - Implementation details
- Experiments
- Summary

Overview

LatentOps



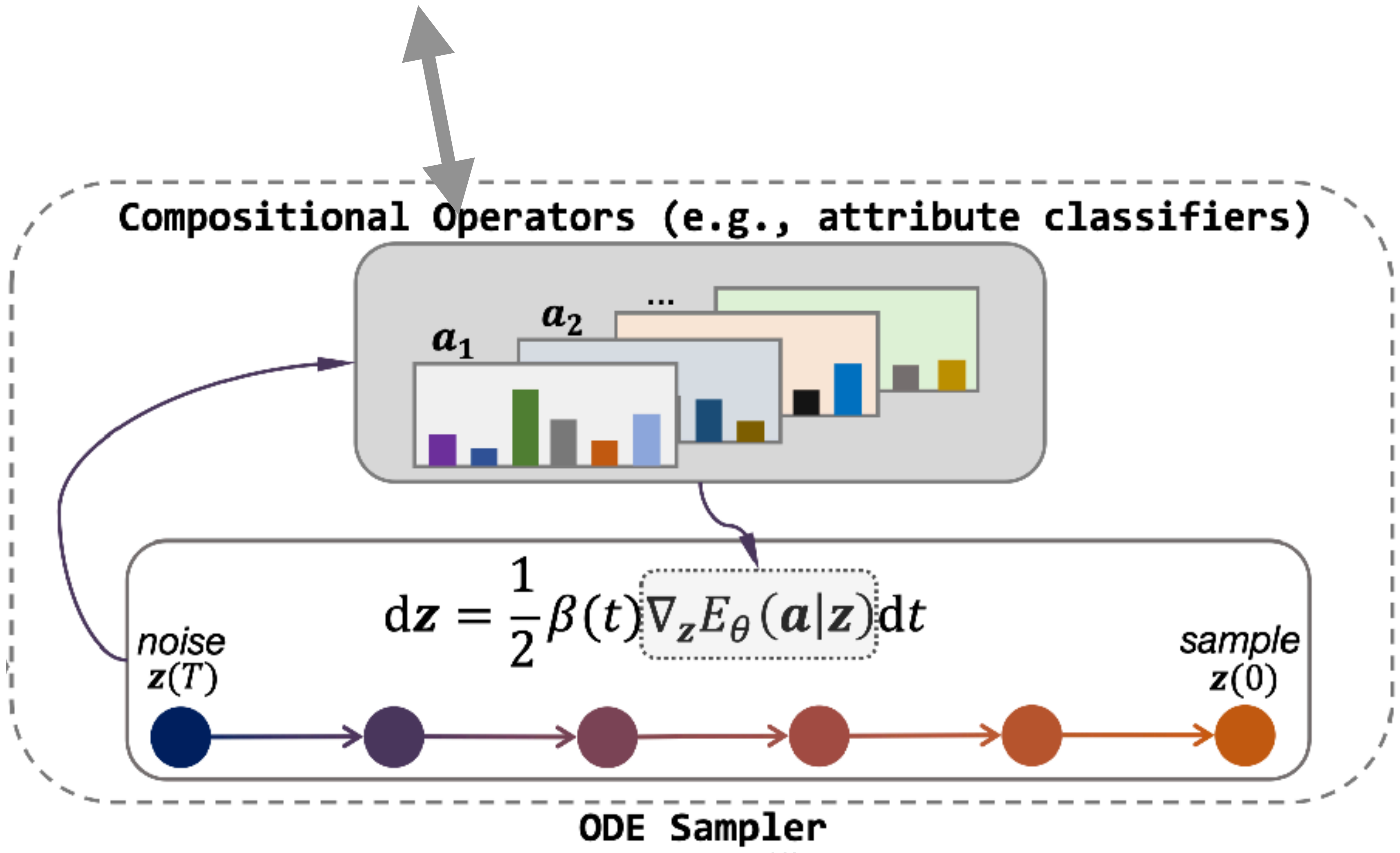
Variational Auto-encoder based on PLMs

Energy-based Model on the latent space

Composable Latent-Space EBMs

- Goal: Formulate a latent-space EBM s.t. one can easily plug in arbitrary attribute operators.
- For Categorical Attribute: to justify whether the desired attributes are in the latent vector

  — Use attribute classifier $f_i(\mathbf{z})$ for attribute a_i



Composable Latent-Space EBMs

- Sample **z** that contains desired attributes **a**
- Joint distribution: $p(\boldsymbol{z}, \boldsymbol{a}) := p_{\text{prior}}(\boldsymbol{z}) p(\boldsymbol{a}|\boldsymbol{z}) = p_{\text{prior}}(\boldsymbol{z}) \cdot e^{-E(\boldsymbol{a}|\boldsymbol{z})}/Z$
- Properties:
  1. Marginal over **z** = **the VAE prior**, i.e., $\sum_{\boldsymbol{a}} p(\boldsymbol{z}, \boldsymbol{a}) = p_{\text{prior}}(\boldsymbol{z})$ -> high quality text.
  2. The energy function enables the **combination of arbitrary attributes** $E(\boldsymbol{a}|\boldsymbol{z}) = \sum_i \lambda_i E_i(a_i|\boldsymbol{z})$
- E_i is defined as the negative log probability of a_i to make sure the different attribute classifiers have outputs at the same scale for combination

$$E_i(a_i|\boldsymbol{z}) = -f_i(\boldsymbol{z})[a_i] + \log \sum_{a_i'} \exp(f_i(\boldsymbol{z})[a_i']).$$

Efficient Sampling with ODEs

Sample from $p(\mathbf{z}, \mathbf{a})$

- Draw samples from $p(\mathbf{z}, \mathbf{a})$
  - Ordinary Differential Equations (ODEs):

$$\begin{aligned} \mathrm{d}\boldsymbol{z} &= -\frac{1}{2}\beta(t)[\boldsymbol{z} + \nabla_{\boldsymbol{z}} \log p_t(\boldsymbol{z}, \boldsymbol{a})]\mathrm{d}t \\ &= -\frac{1}{2}\beta(t)\left[\boldsymbol{z} + \nabla_{\boldsymbol{z}} \log p_t(\boldsymbol{a}|\boldsymbol{z}) + \nabla_{\boldsymbol{z}} \log p_t(\boldsymbol{z})\right]\mathrm{d}t. \end{aligned}$$

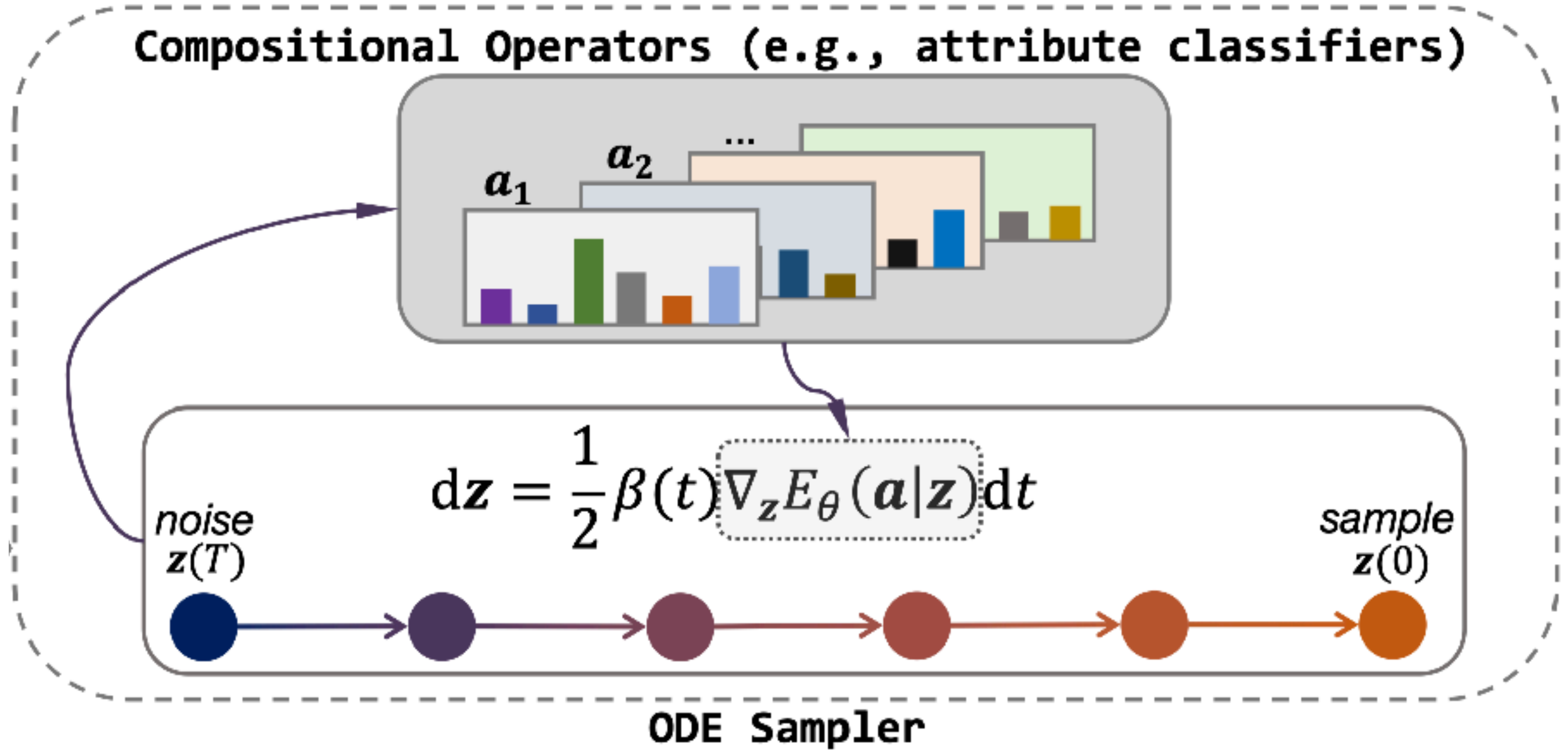
  - $p_0(\mathbf{z}) = p_T(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ —> $p_t(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is **time-invariant.**
  - Classifiers f_i are fixed, and $\mathbf{z}$ is from time-invariant distribution —> $p_t(\mathbf{a} \mid \mathbf{z}) = p(\mathbf{a} \mid \mathbf{z})$ is **time-invariant**

$$\begin{aligned} \mathrm{d}\boldsymbol{z} &= -\frac{1}{2}\beta(t)[\boldsymbol{z} - \nabla_{\boldsymbol{z}} E(\boldsymbol{a}|\boldsymbol{z}) - \frac{1}{2}\nabla_{\boldsymbol{z}}||\boldsymbol{z}||_2^2]\mathrm{d}t \\ &= \frac{1}{2}\beta(t)\sum_{i=1}^{n} \nabla_{\boldsymbol{z}} E(a_i|\boldsymbol{z})\mathrm{d}t. \end{aligned}$$

Summary

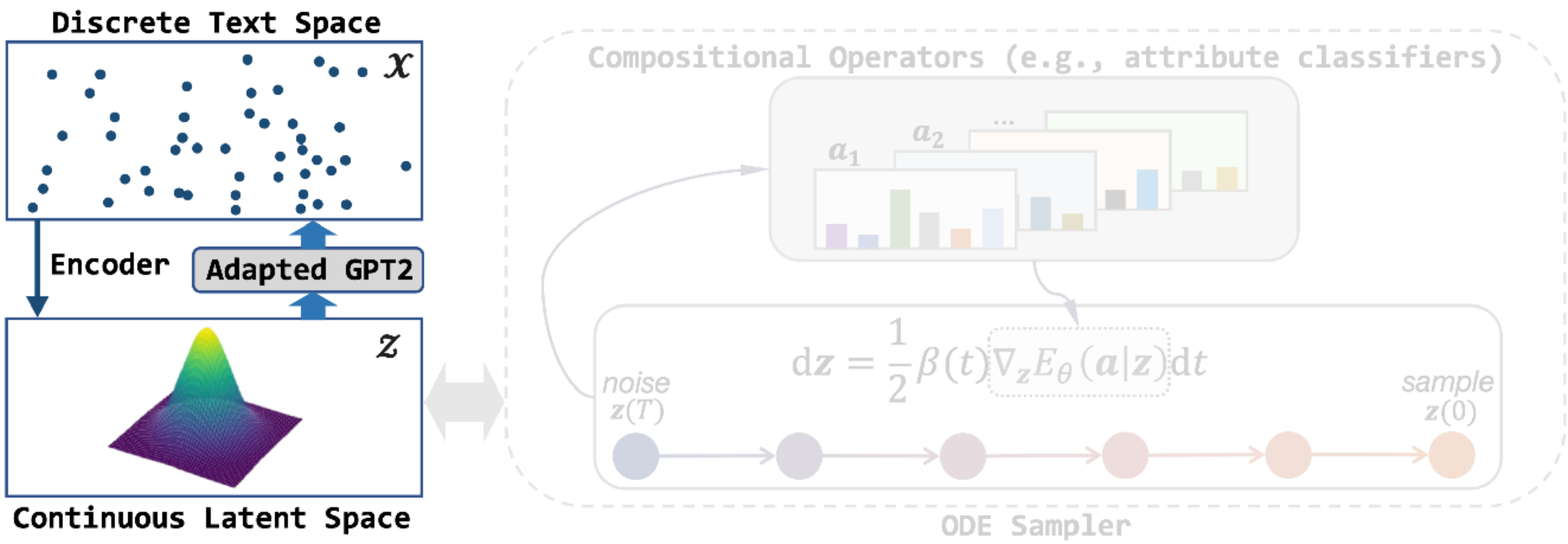
Composable Latent-Space EBMs & Efficient Sampling with ODEs

- Given a text latent space, all we need:
    - Train the attribute classifiers $f_i(\cdot)$ for a_i
    - Sample from $p(\mathbf{z}, \mathbf{a})$ by solve the ODE.
    - Different classifiers can be freely combined.



Latent Model

VAE



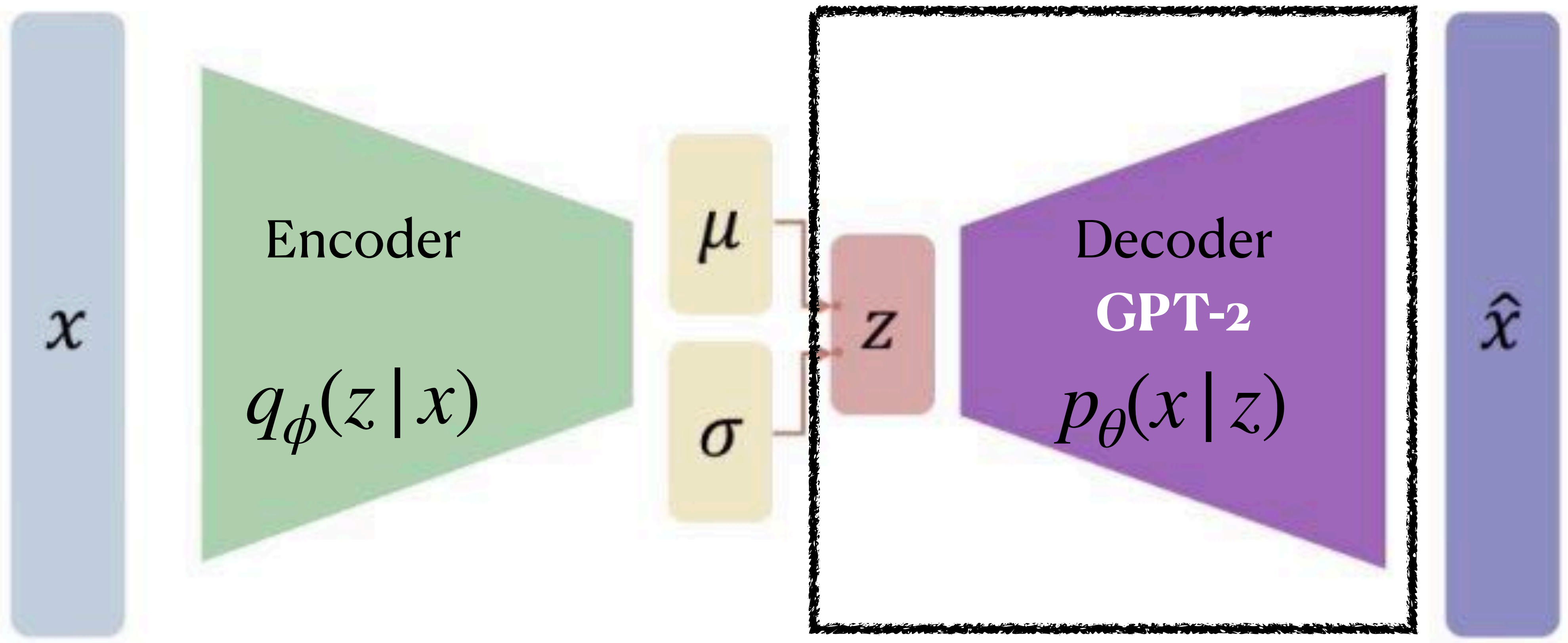
Adapting Pretrained LMs for Latent Space

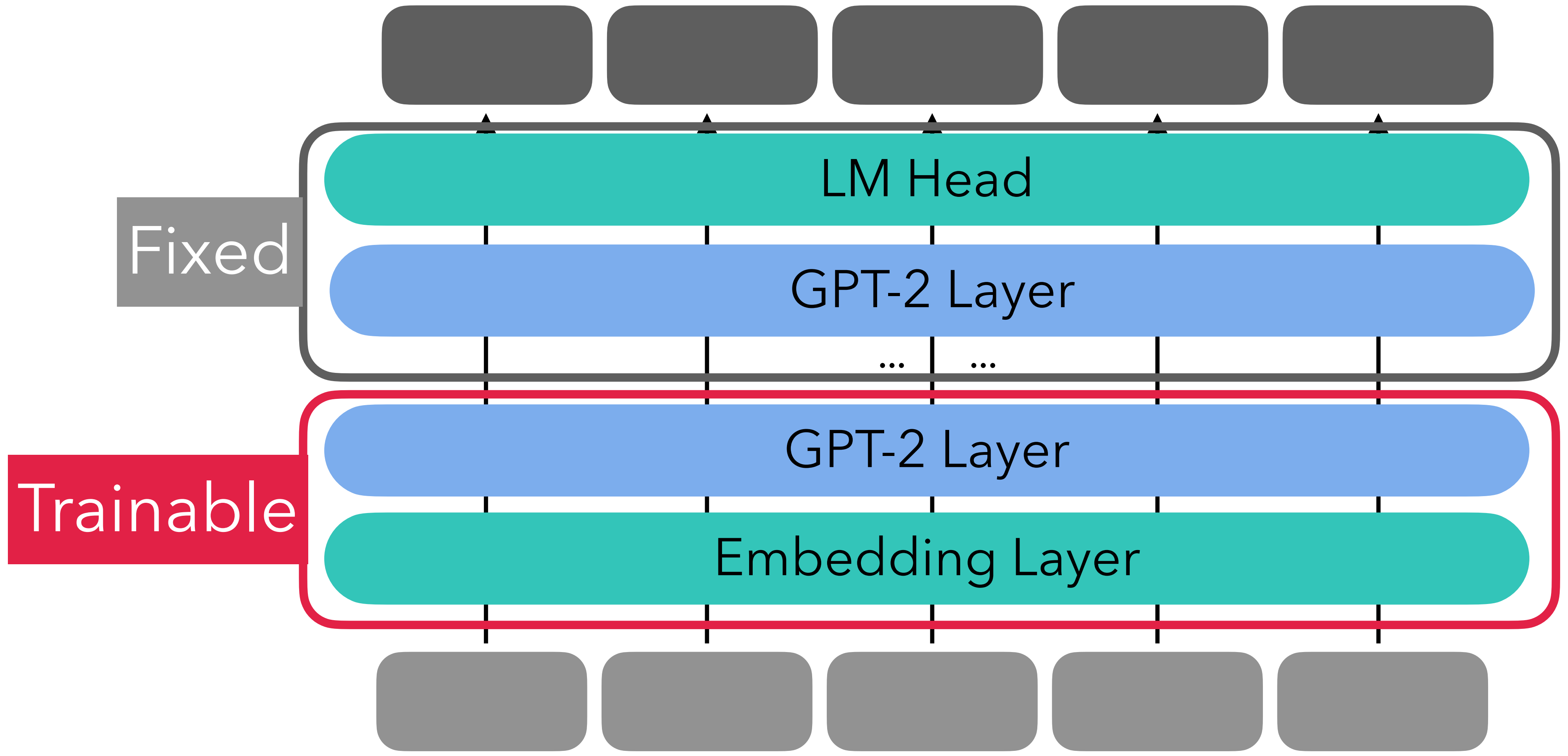
Variational Auto-Encoder

Decoder:

Equip PLMs (e.g., GPT-2) with the latent space through **parameter-efficient adaptation**.

- Update ***a small portion*** of the LM parameters
- Keep the LM's ability to generate fluent and coherent text
- Adopt simple MLP layers that pass the latent vector to the LM (Embedding and Attention)





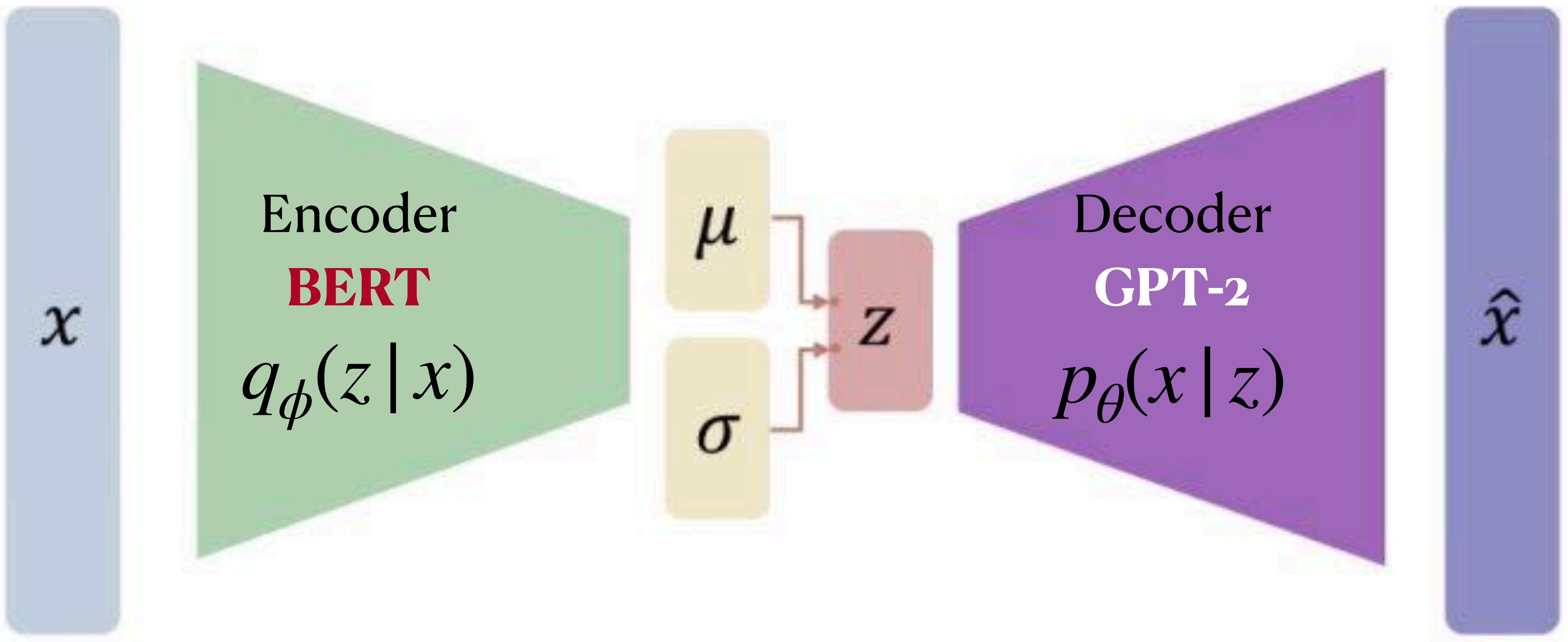
Adapting Pretrained LMs for Latent Space

Variational Auto-Encoder

Encoder:

We use a BERT-small, and fine-tune it in the VAE framework.

The tuned encoder can be used to produce the initial z values in the ODE sampler for text editing.



Implementation Details

How to acquire attribute classifiers

- Train attribute classifiers on the frozen latent space.
- Map text **x** into latent space -> training pairs (**z**, **a**)
- Since the classifier is built on the semantic latent space, it can be trained efficiently with only **a small number of examples** (e.g., 200 per class)
  - Don't require large amount of labeled data

Implementation Details

Initialization of ODE sampling

- For generating new text:

  Initialize $\mathbf{z}(T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

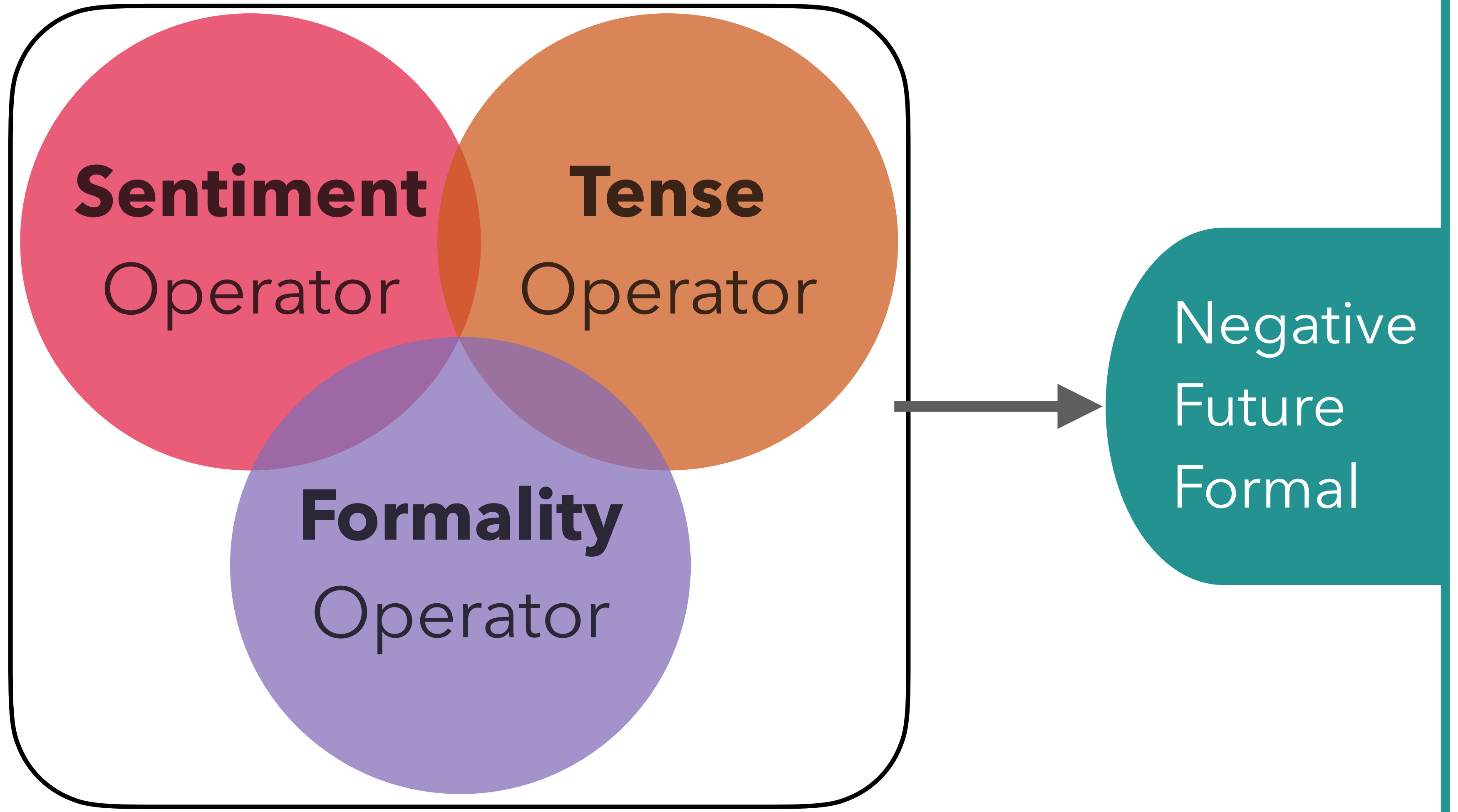
- For text editing:
  - The main content should be preserved.
  - Initialize $\mathbf{z}(T) \sim q_\phi(\mathbf{z} \,|\, \mathbf{x})$ (the latent vector of the given text by encoder)
  - $\mathbf{z}(0)$ preserves content and obtains desired attributes well.

Outline

- Problem Statement
- Background
- Method
  - Composable Latent-Space EBMs
  - Efficient Sampling with ODEs
  - Adapting Pretrained LMs for Latent Space
  - Implementation details
- **Experiments**
- Summary

Experiments

Generation with Compositional Attributes



Negative + Future + Formal

GPT2-FT:
- i will not be back.
- would not recommend this location to anyone. [No Subject]
- would not recommend them for any jewelry or service. [No Subject]
- if i could give this place zero stars, i would.

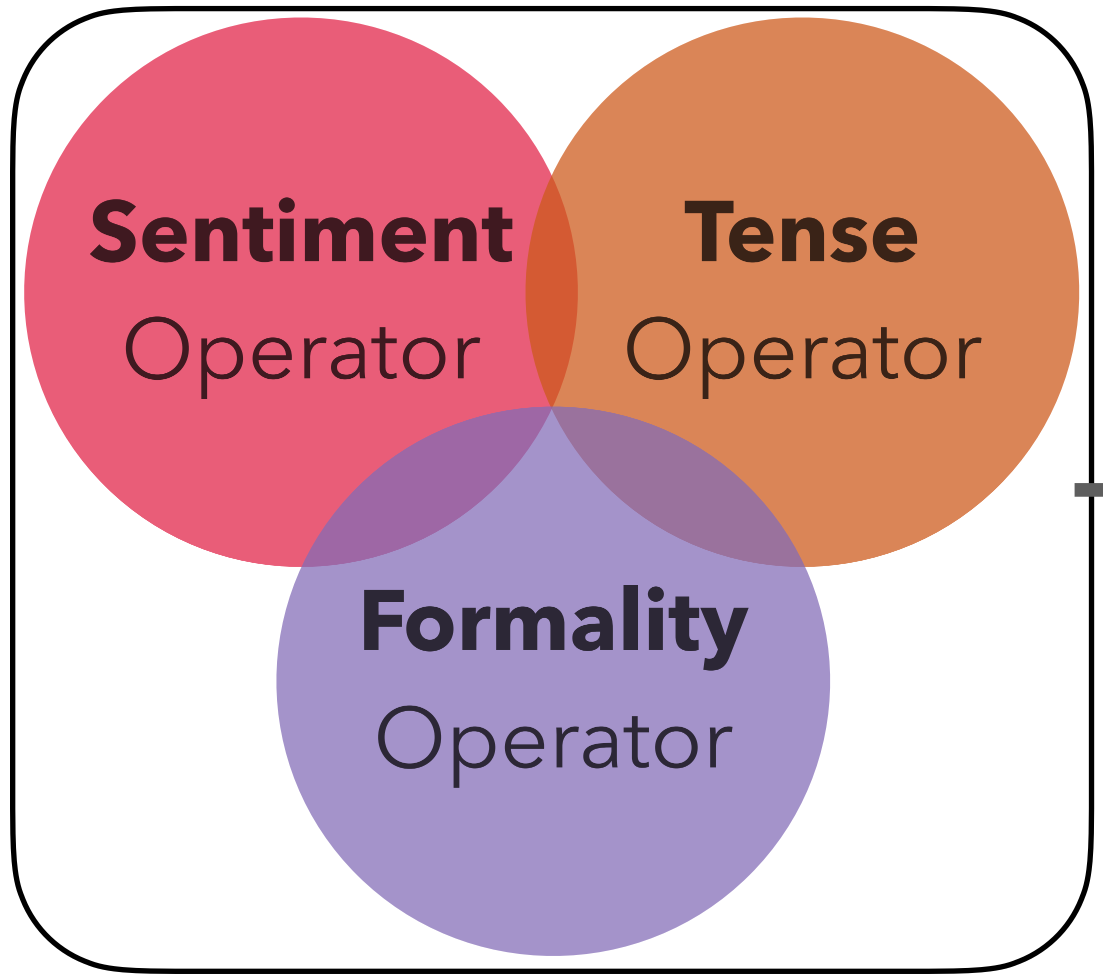
PPLM:
- i could not recommend them at all.
- i could not believe this was not good!
- this was a big deal, because the food was great.
- i could not recommend them.

FUDGE:
- not a great pizza to get a great pie! [No Tense]
- however, this place is pretty good.
- i have never seen anything like these.
- will definitely return. [No Subject]

Ours:
- i would not believe them to stay .
- i will never be back .
- i would not recommend her to anyone in the network .
- they will not think to contact me for any reason .

Experiments

Generation with Compositional Attributes



All combination

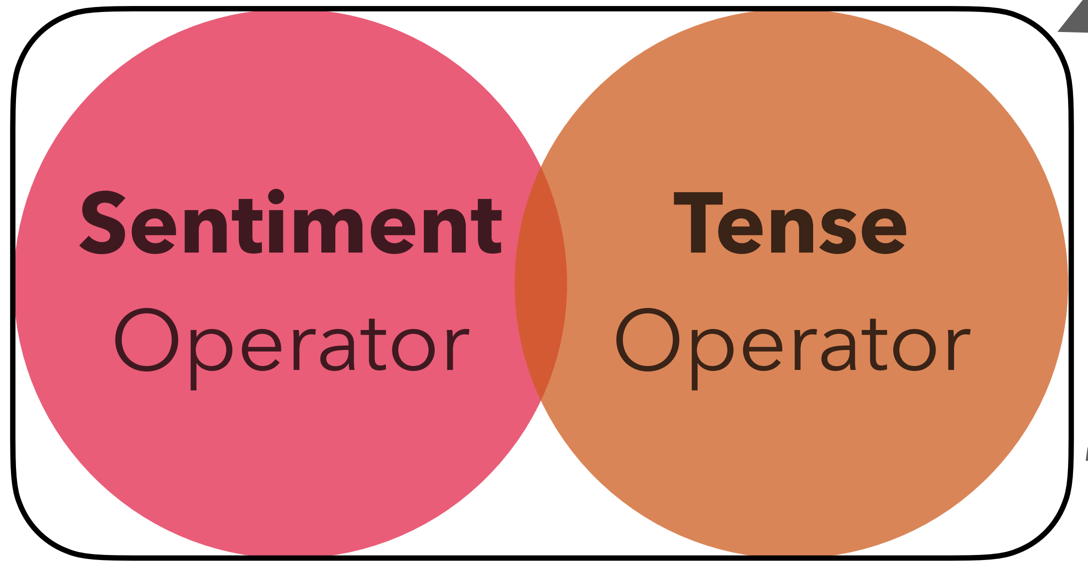
Attributes	Methods	Accuracy↑				Fluency↓	Diversity↓
		S	T	F	G-M	PPL	self-BLEU
Sentiment	GPT2-FT	0.98	-	-	0.98	10.6	23.8
	PPLM	0.86	-	-	0.86	11.8	31.0
	FUDGE	0.77	-	-	0.77	**10.3**	27.2
	Ours	**0.99**	-	-	**0.99**	30.4	**13.0**
Sentiment +Tense	GPT2-FT	0.98	0.95	-	0.969	9.0	36.8
	PPLM	0.81	0.59	-	0.677	15.7	28.7
	FUDGE	0.67	0.63	-	0.565	**11.0**	35.9
	Ours	**0.98**	**0.93**	-	**0.951**	25.2	**19.7**
Sentiment +Tense +Formality	GPT2-FT	0.97	0.92	0.87	0.919	10.3	36.8
	PPLM	0.82	0.57	0.56	0.598	17.5	30.5
	FUDGE	0.67	0.64	0.62	0.556	**11.5**	35.9
	Ours	**0.97**	**0.92**	**0.93**	**0.937**	25.8	**21.1**

Time for generating 150 samples

Methods	PPLM	FUDGE	Ours
Time (s)	3182 (578×)	36.1 (6.6×)	5.5 (1×)

Examples

Text Editing with Compositional Attributes



Source	this place is a terrible place to live !
Human	this place is a great place to live !
FUDGE	great place to live!
+ Past	great food and terrible service! [No Tense]
+ Present	great place to live! [No Tense]
+ Future	great place to live! [No Tense]
Ours	this place is a great place to live !
+ Past	this place was a great place to live !
+ Present	this place is a great place to live !
+ Future	this place would have a great place to live !

Experiments

Text Editing

Methods	Accuracy↑	Content↑			Fluency↓	#Params	#Data
	Sentiment	iBL	rBL	CTC	PPL		
Source	0.27	100	31.4	0.500	15.9	-	-
Human	0.82	31.9	100	0.463	24.5	-	-
B-GST	0.81	31.8	16.3	0.473	39.5	111M	Full-data
STrans	0.91	53.2	24.5	0.469	41.0	17M	
DiRR	**0.96**	**61.5**	**29.8**	**0.480**	23.9	1.5B	
T&G	0.88	47.6	21.8	0.466	24.3	63M	
FGST	0.90	13.2	7.6	0.450	**9.3**	26M	
FUDGE	0.40	57.0	18.0	0.456	39.3	16.4M	**Few-shot**
Ours	0.95	54.0	24.3	0.474	25.9	**3.7K**	
Source	0.14	100	49.4	0.425	26.4	-	-
Human	0.52	49.7	100	0.422	47.2	-	-
B-GST	0.62	52.3	28.5	0.425	27.7	111M	Full-data
DiRR	0.60	68.7	**38.2**	0.424	32.5	1.5B	
T&G	0.65	68.6	35.4	0.423	40.9	63M	
FGST	**0.83**	21.9	14.0	**0.427**	**13.6**	26M	
FUDGE	0.20	**70.5**	35.1	0.415	49.5	16.4M	**Few-shot**
Ours	0.72	53.3	28.1	0.423	44.1	**3.7K**	

Methods	Accuracy↑		Content↑		Fluency↓
	Sentiment	Tense	iBL	CTC	PPL
FUDGE	0.36	0.56	**56.5**	0.450	**17.3**
Ours	**0.95**	**0.95**	37.1	**0.465**	30.1

Summary

LatentOps

- A new **efficient** approach that performs **composable** control operations in the **compact** and **continuous latent space** of text.
- Permits plugging in **arbitrary operators** to form an **energy-based distribution** on the **low-dimensional** latent space.
- We develop an efficient sampler based on **ODEs** to draw latent vector samples that **bear the desired attributes**.
- We connect the latent space to pretrained LMs (e.g., GPT-2) by efficiently **adapting a small subset of the LM parameters** in a variational auto-encoding (**VAE**) manner.

Some Observations

- The demand of capacity of VAE encoder is not great.
- The generation quality mainly depends on the decoder capacity.
- VAE is not a perfect choice as the latent model.
  - Tradeoff between reconstruction and generation
  - Gap between posterior and prior.

Thanks!

Discussions and collaborations are welcome!

Contact me:

guangyiliu@link.cuhk.edu.cn

Our Group @UCSD:

http://zhiting.ucsd.edu/

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

UC San Diego