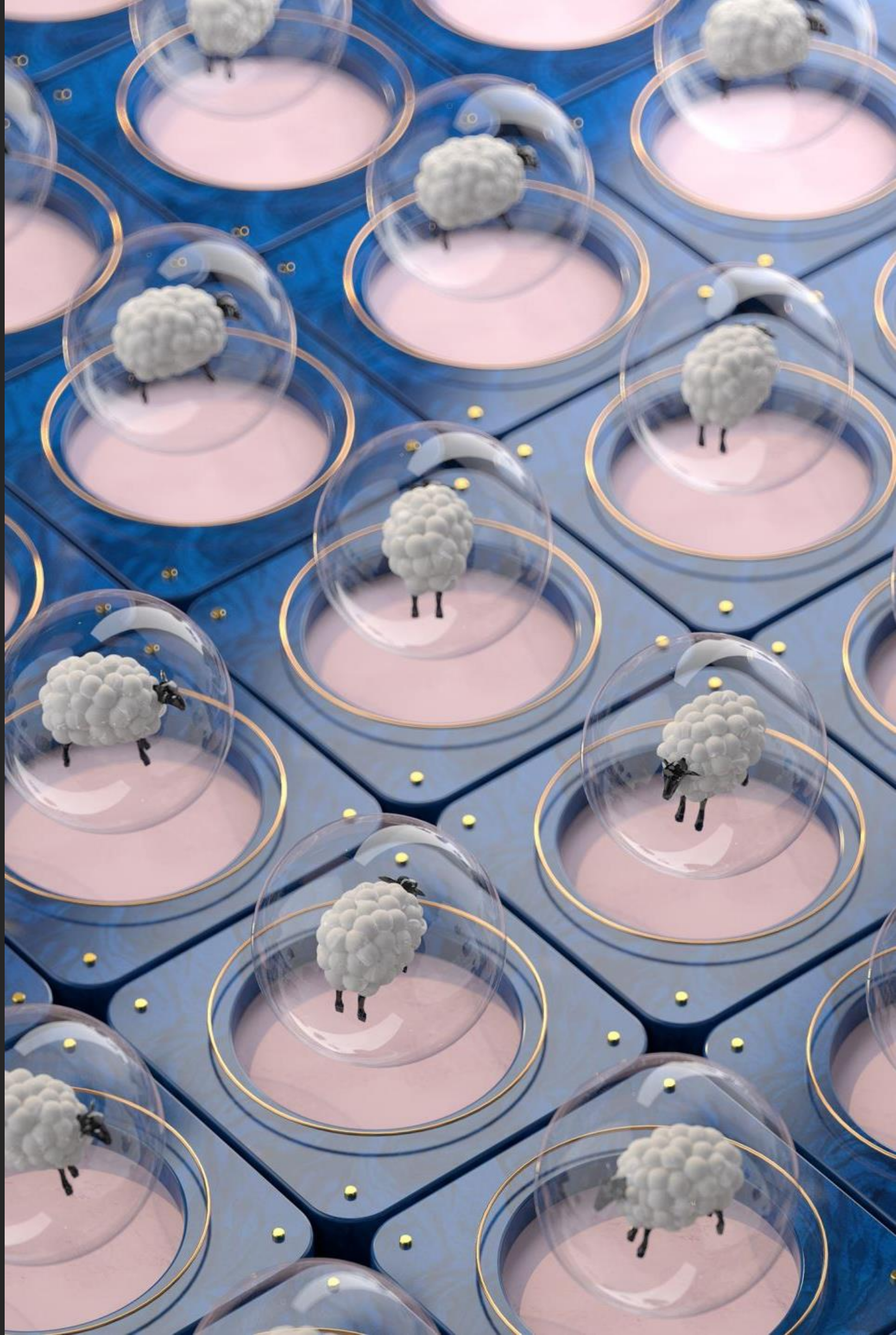
Using Diffusion Models to Generate Counterfactual Objects

TEAM WHITE – CS 726 PROJECT

PROJECT GUIDE: SUNITA SARAWAGI

- SHIV KIRAN BAGATHI 200050019
- SAI KIRAN 200050023
- TANUJA 200050029
- SASANK 200050045



Problem Statement

TASK OF GENERATING COUNTERFACTUAL DATA FROM OBSERVATIONAL IMAGING DATA

Problems we face while generating Counterfactuals

- Generating counterfactuals is difficult because it requires predicting what would have happened if certain aspects of the past had been different.
- This is challenging because it requires modeling the complex causal relationships between variables and accounting for all the confounding factors that may have influenced the outcome.
- The data available may not be sufficient to accurately estimate the counterfactual outcome when using neural networks.

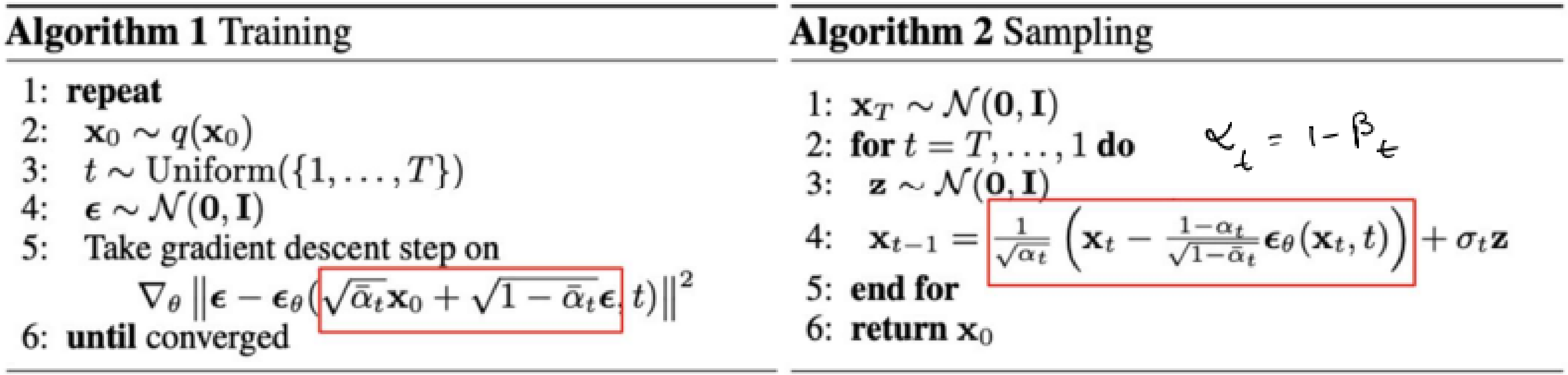
Why Diffusion Models ?

- Stochasticity (randomness) in the diffusion process, which is used in modeling causal relationships between variables, can be helpful in building causal models that are aware of and account for uncertainty. We can also quantify the uncertainty.
- The iterative sampling can be naturally extended for applying interventions. Modifying initial condition and score function can simulate the interventions.

Short Intro to Diffusion Models

$$p\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{\alpha_t} \mathbf{x}_0,\left(1-\alpha_t\right) \mathrm{I}\right), \text{ where } \alpha_t:=\prod_{j=0}^t\left(1-\beta_j\right)$$

$$\theta^*=\underset{\theta}{\arg \min } \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0, \mathrm{I})}\left[\left(1-\alpha_t\right)\left\|\boldsymbol{\epsilon}_\theta\left(\sqrt{\alpha_t} \mathbf{x}_0+\sqrt{1-\alpha_t} \epsilon, t\right)-\epsilon\right\|_2^2\right] .$$



Algorithm 1 Training

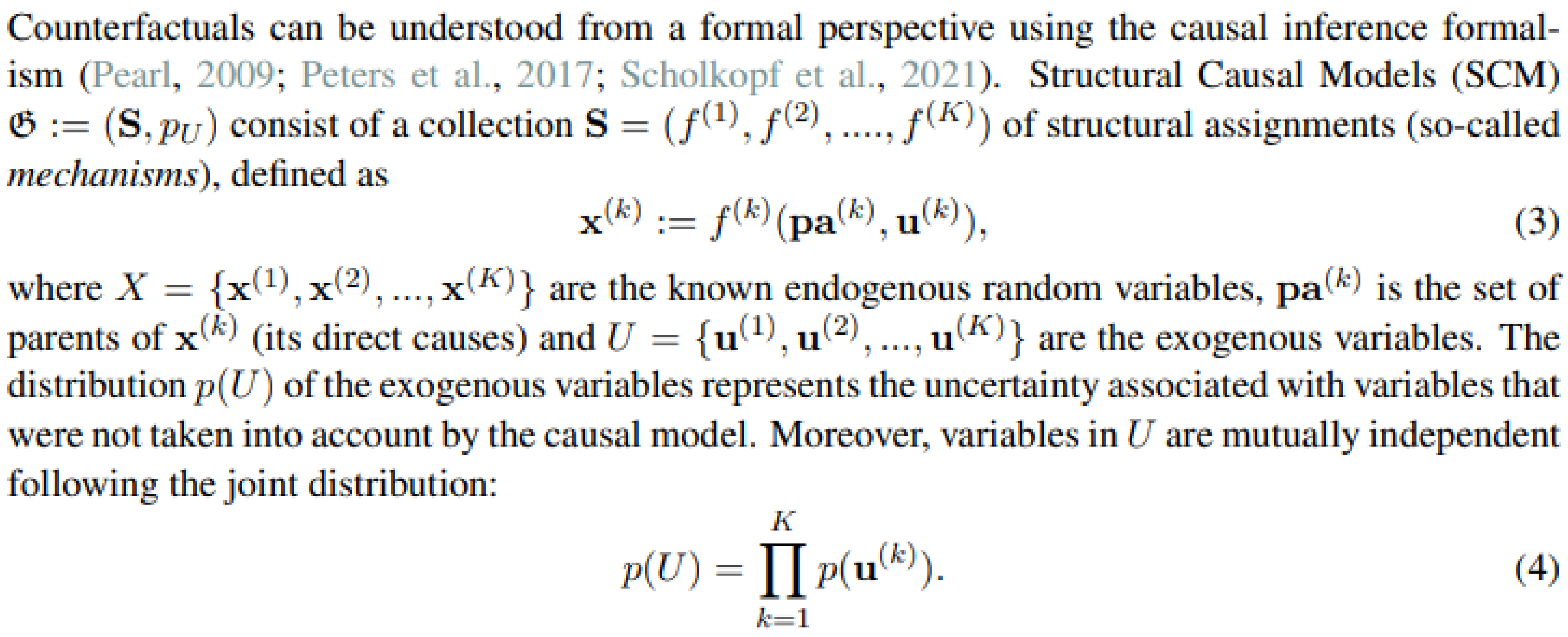
```
repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
      $\nabla_\theta \left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boxed{\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}}, t\right)\right\|^2$
until converged
```

Algorithm 2 Sampling

```
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t=T, \ldots, 1$ do
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_{t-1}=\boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right)}+\sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$
```

$\alpha_t = 1-\beta_t$

Intro to Structural Causal Model



Counterfactuals can be understood from a formal perspective using the causal inference formalism (Pearl, 2009; Peters et al., 2017; Scholkopf et al., 2021). Structural Causal Models (SCM) $\mathfrak{G} := (\mathbf{S}, p_U)$ consist of a collection $\mathbf{S} = (f^{(1)}, f^{(2)}, ..., f^{(K)})$ of structural assignments (so-called *mechanisms*), defined as

$$\mathbf{x}^{(k)} := f^{(k)}(\mathbf{pa}^{(k)}, \mathbf{u}^{(k)}), \tag{3}$$

where $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(K)}\}$ are the known endogenous random variables, $\mathbf{pa}^{(k)}$ is the set of parents of $\mathbf{x}^{(k)}$ (its direct causes) and $U = \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, ..., \mathbf{u}^{(K)}\}$ are the exogenous variables. The distribution $p(U)$ of the exogenous variables represents the uncertainty associated with variables that were not taken into account by the causal model. Moreover, variables in U are mutually independent following the joint distribution:

$$p(U) = \prod_{k=1}^{K} p(\mathbf{u}^{(k)}). \tag{4}$$

Some Terms related to Causal Models

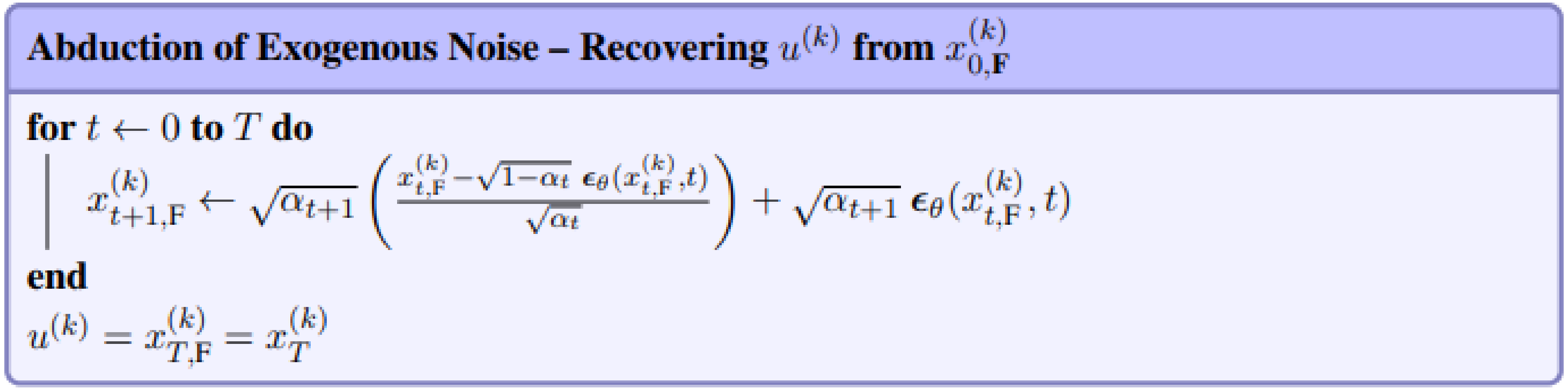
Endogenous variables : Variables that are influenced by other variables within a system, and their values are determined by the relationships between these variables.

Causal Variables : Variables that cause changes in other variables in the system.

Exogenous Variables : Variables that are not influenced by other variables within the system and are assumed to be independent of the system. Ex: Like lighting and contrast of image.

Modification of Forward Step

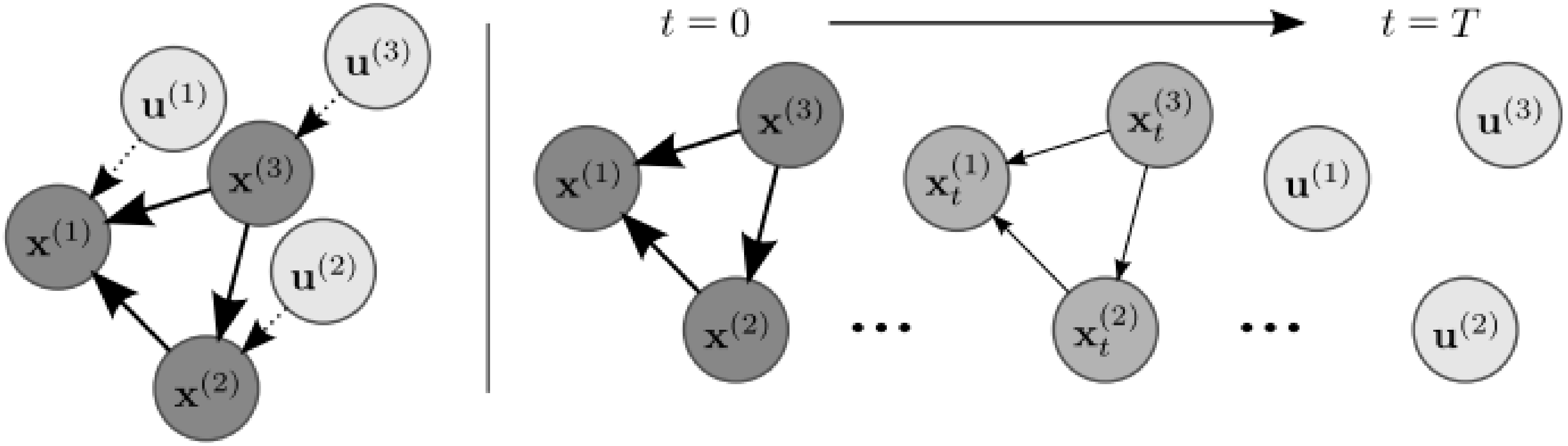
Recovering of Exogenous variables from input image using trained diffusion model on input data.



Abduction of Exogenous Noise – Recovering $u^{(k)}$ from $x_{0,\mathbf{F}}^{(k)}$

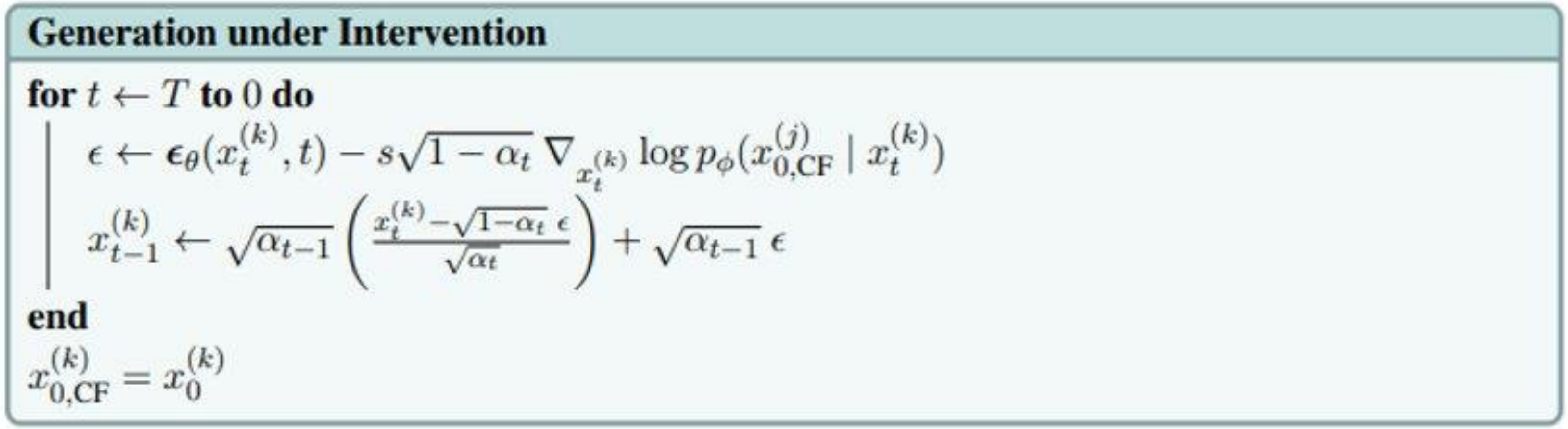
```
for $t \leftarrow 0$ to $T$ do
    $x_{t+1,\mathrm{F}}^{(k)} \leftarrow \sqrt{\alpha_{t+1}} \left( \frac{x_{t,\mathrm{F}}^{(k)} - \sqrt{1-\alpha_t}\ \epsilon_\theta(x_{t,\mathrm{F}}^{(k)}, t)}{\sqrt{\alpha_t}} \right) + \sqrt{\alpha_{t+1}}\ \epsilon_\theta(x_{t,\mathrm{F}}^{(k)}, t)$
end
$u^{(k)} = x_{T,\mathrm{F}}^{(k)} = x_T^{(k)}$
```

Effect of Forward Step on SCM



Modification on Reverse Process

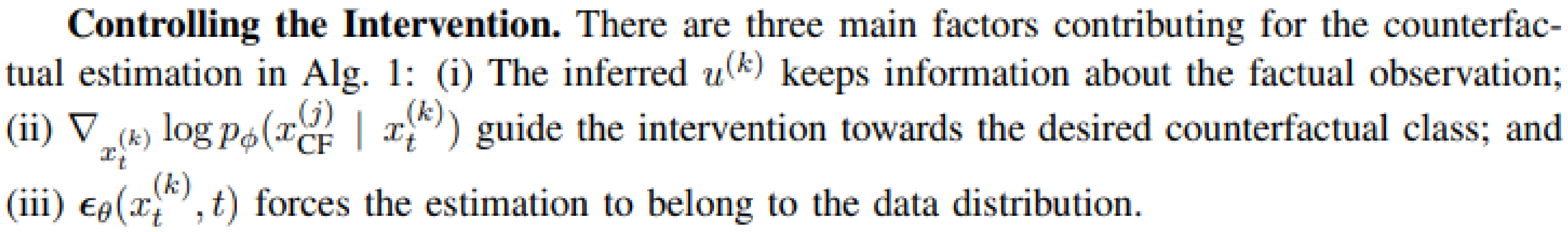
Applying Anti-Causal Predictors to error, generate counterfactuals.



Generation under Intervention

```
for $t \leftarrow T$ to $0$ do
    $\epsilon \leftarrow \boldsymbol{\epsilon}_\theta(x_t^{(k)}, t) - s\sqrt{1 - \alpha_t}\, \nabla_{x_t^{(k)}} \log p_\phi(x_{0,\mathrm{CF}}^{(j)} \mid x_t^{(k)})$
    $x_{t-1}^{(k)} \leftarrow \sqrt{\alpha_{t-1}} \left( \frac{x_t^{(k)} - \sqrt{1-\alpha_t}\, \epsilon}{\sqrt{\alpha_t}} \right) + \sqrt{\alpha_{t-1}}\, \epsilon$
end
$x_{0,\mathrm{CF}}^{(k)} = x_0^{(k)}$
```

Controlling the Intervention



Controlling the Intervention. There are three main factors contributing for the counterfactual estimation in Alg. 1: (i) The inferred $u^{(k)}$ keeps information about the factual observation; (ii) $\nabla_{x_t^{(k)}} \log p_\phi(x_{\text{CF}}^{(j)} \mid x_t^{(k)})$ guide the intervention towards the desired counterfactual class; and (iii) $\boldsymbol{\epsilon}_\theta(x_t^{(k)}, t)$ forces the estimation to belong to the data distribution.

Implementation of the Model

We performed the given model to generate counterfactuals on MNIST dataset.

The paper used Spaced diffusion which is a variant of the diffusion process used in diffusion models, which involves sequentially applying a series of diffusion steps with intermediate observations, instead of using a single long diffusion process.

We used linear beta scheduling for the forward process.

Unnet for training of error as both input and output shapes of images are same.

Implementation of Model

The Unet is utilized to train the score model for predicting noise.

A Half Unet is employed to determine the Anti-Causal Predictor.

During the generation of counterfactuals, the Unet is conditioned on the Anti-Causal Predictor.

References

- This model is heavily based on Diffusion Causal Models for Counterfactual Estimation by Pedro Sanchez, Sotirios A. Tsaftaris https://arxiv.org/pdf/2202.10166.pdf
- Denoising Diffusion Probabilistic Models Jonathan Ho, Ajay Jain, Pieter Abbeel https://arxiv.org/pdf/2006.11239
- A great intuition on Diffusion Models by Computerphile https://www.youtube.com/watch?v=1CIpzeNxIhU

Thank You