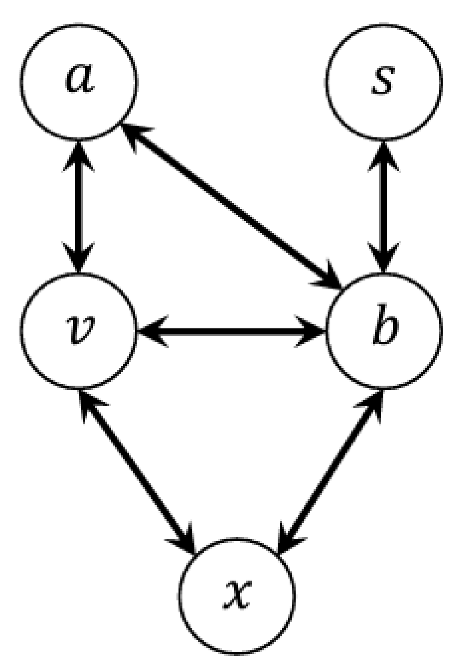
Causal Diffusion Autoencoders: Toward Counterfactual Generation via Diffusion Probabilistic Models

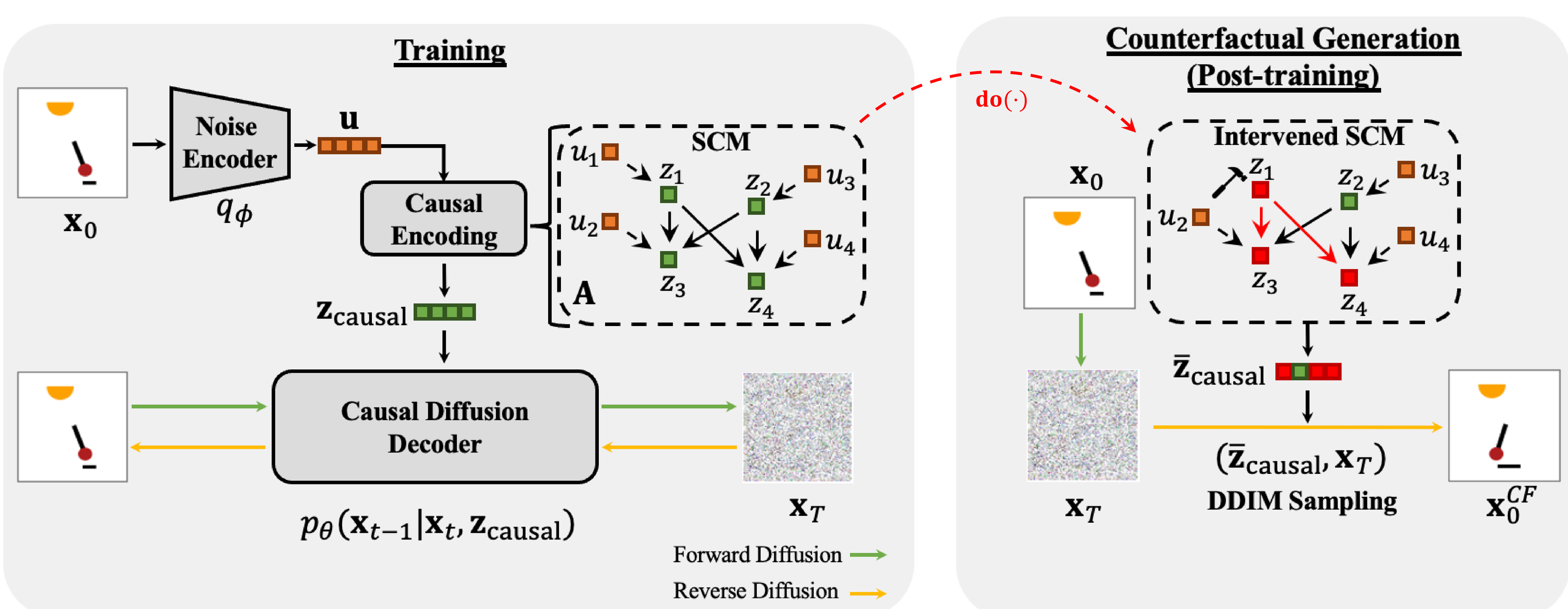
Aneesh Komanduri[1], Chen Zhao[2], Feng Chen[3], and Xintao Wu[1]
[1] University of Arkansas, Fayetteville, AR, USA
[2] Baylor University, Waco, TX, USA
[3] University of Texas at Dallas, Richardson, TX, USA

Motivation

- Diffusion models have shown impressive capability for image generation, but **counterfactual generation** (i.e., generating hypothetical scenarios consistent with a causal graph) has not been explored much.
- **Why counterfactual generation?** In domains such as healthcare, modeling causal variables that underlie the image generation process can lead to the ability to generate hypothetical scenarios that help with **reducing data collection costs** and **planning treatments**.
- **Motivating Example**: Brain MRI scan where age (a) causes brain volume (b)
- *How can we utilize causality in diffusion models to improve counterfactual generation capabilities?*



Methodology



- We propose **CausalDiffAE**, a diffusion-based causal representation learning framework to enable counterfactual generation.
- **General Strategy**:
  - **Model latent causal mechanisms:** Encode image to a noise representation $\mathbf{u}$ and map to causal variables $\mathbf{z}_{\text{causal}}$ via neural networks

$$z_i = f_i(z_{\mathbf{pa}_i}, u_i)$$

  - **Representation-conditioned diffusion model:** Condition the diffusion model on causal variables $\mathbf{z}_{\text{causal}}$ extracted from high dimensional input data along with stochastic noise $\mathbf{x}_T$

$$\mathcal{L}_{\text{CausalDiffAE}} = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0, \epsilon_t}\left[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{causal}}) - \epsilon_t\|_2^2\right] + \gamma\Big\{\mathcal{D}_{KL}(q_\phi(\mathbf{z}_{\text{causal}}|\mathbf{x}_0, \mathbf{y})\|p(\mathbf{z}_{\text{causal}}|\mathbf{y})) + \mathcal{D}_{KL}(q_\phi(\mathbf{u}|\mathbf{x}_0)\|\mathcal{N}(\mathbf{0}, \mathbf{I}))\Big\}$$

$$p(\mathbf{z}_{\text{causal}}|\mathbf{y}) = \prod_{i=1}^{n} p(z_i|y_i) = \prod_{i=1}^{n} \mathcal{N}(z_i; \mu_\nu(y_i), \sigma_\nu^2(y_i)\mathbf{I})$$

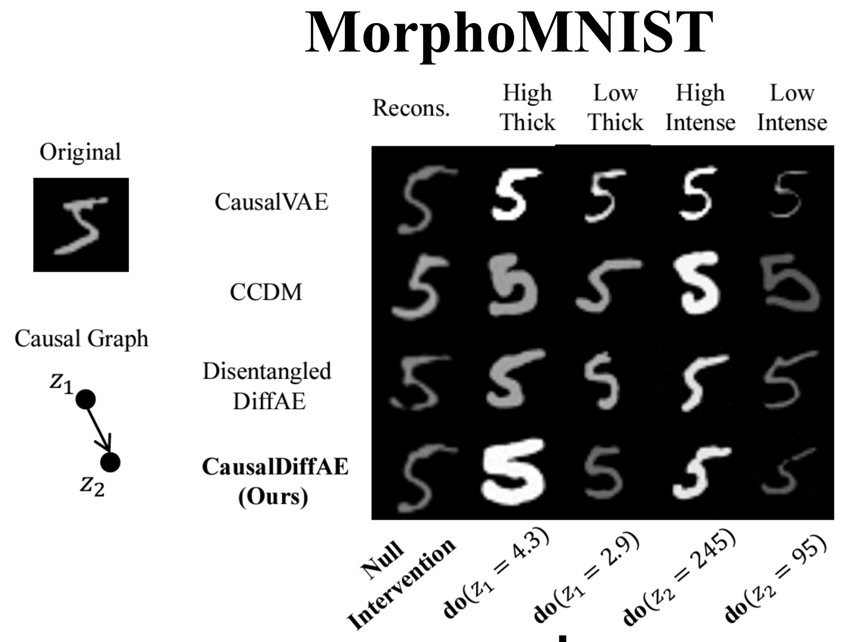
  - **Generate Counterfactuals:** Manipulate latent variable and deterministically decode to counterfactual with intervened latents $\bar{\mathbf{z}}_{\text{causal}}$

$$\mathbf{x}_{t-1}^{CF} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t^{CF} - \sqrt{1-\alpha_t}\,\epsilon_\theta(\mathbf{x}_t^{CF}, t, \bar{\mathbf{z}}_{\text{causal}})}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(\mathbf{x}_t^{CF}, t, \bar{\mathbf{z}}_{\text{causal}})$$

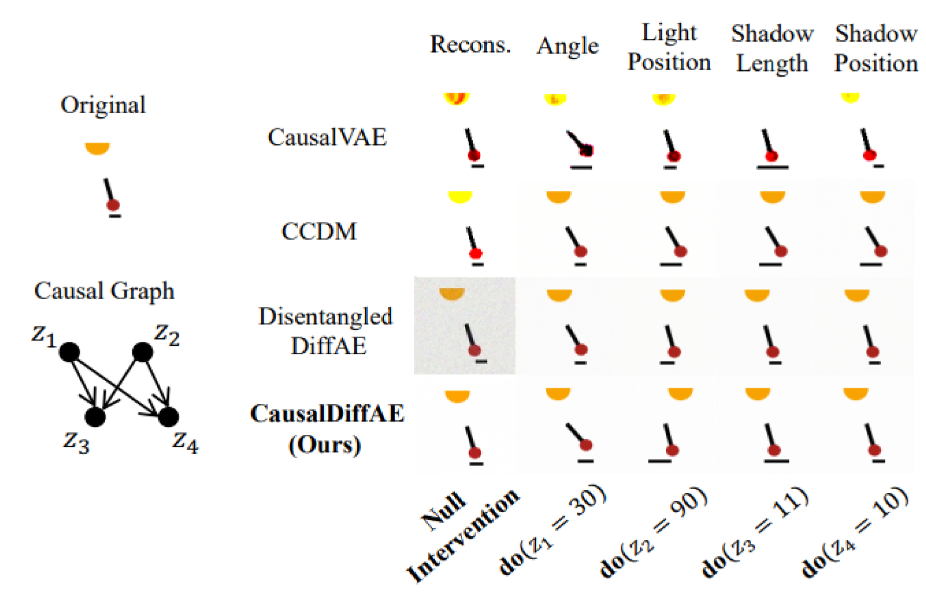
Experimental Evaluation

Counterfactual Generation

MorphoMNIST

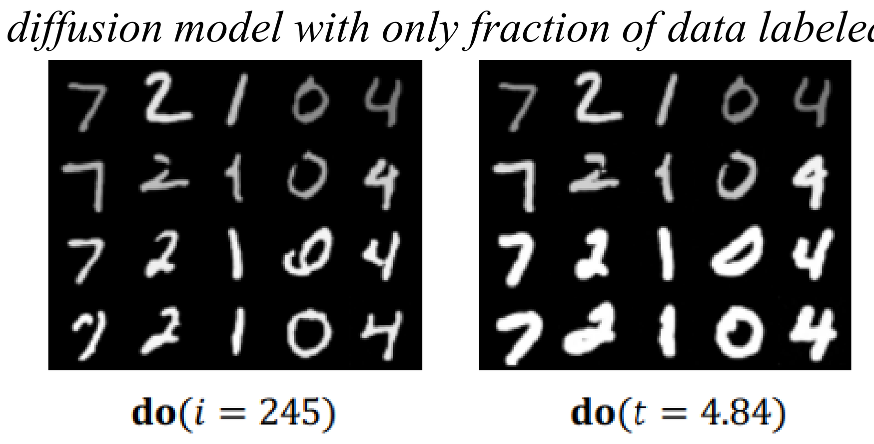


Pendulum



Weak Supervision Case-Study

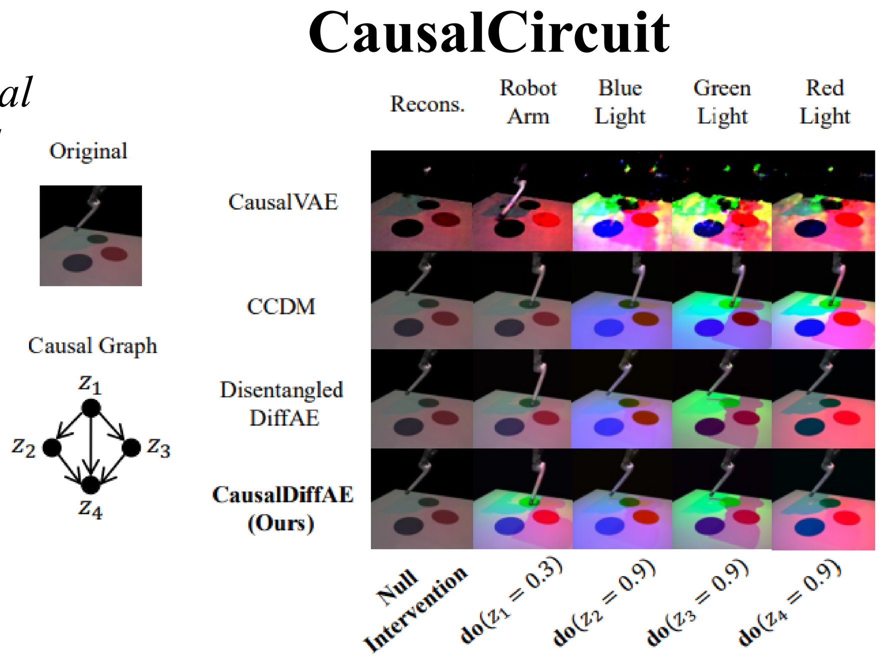
Strategy: *Jointly train conditional and unconditional diffusion model with only fraction of data labeled*



$$\bar{\epsilon}_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}}_{\text{causal}}) = \underbrace{\omega\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}}_{\text{causal}})}_{\text{causal conditional model}} + \underbrace{(1-\omega)\epsilon_\theta(\mathbf{x}_t, t)}_{\text{unconditional model}}$$

Enables granular control over generated counterfactuals as we change ω!

CausalCircuit



Disentanglement

- DCI disentanglement score to evaluate the degree of non-overlapping in learned causal factors
- **High disentanglement** in causal diffusion-based objective implies controllability of learned variables

Dataset	Model	DCI ↑
MorphoMNIST	CausalVAE	0.784 ± 0.01
	DiffAE	0.358 ± 0.01
	CausalDiffAE	$\mathbf{0.993 \pm 0.01}$
Pendulum	CausalVAE	0.885 ± 0.01
	DiffAE	0.353 ± 0.01
	CausalDiffAE	$\mathbf{0.999 \pm 0.01}$
CausalCircuit	CausalVAE	0.8860 ± 0.01
	DiffAE	0.353 ± 0.01
	CausalDiffAE	$\mathbf{0.999 \pm 0.01}$

Effectiveness

- The **effectiveness** metric evaluates how accurate the counterfactual is with respect to the true counterfactual
  - (1) Train **anti-causal classifiers** for each causal variable given a training dataset
  - (2) Generate counterfactual via generative model, feed into trained anti-causal classifier and compare prediction to ground-truth counterfactual label values
- **CausalDiffAE** generated counterfactuals yield **low MAE** for nearly all predicted causal factors upon interventions on learned causal factors

MorphoMNIST

Factor	Model	Intervention	
		$\mathbf{do}(t)$	$\mathbf{do}(i)$
Thickness (t)	CausalVAE	3.763 ± 0.01	4.645 ± 0.01
	DisDiffAE	$\mathbf{0.377 \pm 0.02}$	0.326 ± 0.02
	CausalDiffAE	0.392 ± 0.02	$\mathbf{0.309 \pm 0.02}$
Intensity (i)	CausalVAE	13.233 ± 0.01	15.087 ± 0.01
	DisDiffAE	0.794 ± 0.02	0.262 ± 0.02
	CausalDiffAE	$\mathbf{0.503 \pm 0.01}$	$\mathbf{0.256 \pm 0.01}$

Pendulum

Factor	Model	Intervention			
		$\mathbf{do}(a)$	$\mathbf{do}(lp)$	$\mathbf{do}(sl)$	$\mathbf{do}(sp)$
Angle (a)	CausalVAE	24.860	23.030	20.470	11.580
	DisDiffAE	0.668	0.648	0.647	0.647
	CausalDiffAE	**0.297**	**0.132**	**0.031**	**0.034**
LightPos (lp)	CausalVAE	34.200	26.010	35.490	47.060
	DisDiffAE	0.656	0.654	0.630	0.651
	CausalDiffAE	**0.045**	**0.434**	**0.035**	**0.064**
ShadowLen (sl)	CausalVAE	1.946	1.43	2.02	1.72
	DisDiffAE	0.550	0.527	0.560	0.516
	CausalDiffAE	**0.136**	**0.322**	**0.492**	**0.082**
ShadowPos (sp)	CausalVAE	52.52	72.50	57.03	32.78
	DisDiffAE	0.474	0.475	0.479	0.534
	CausalDiffAE	**0.146**	**0.303**	**0.064**	**0.471**

* Standard error is roughly in the range ± 0.01 to ± 0.02 for all averages.

Acknowledgement: This work was supported in part by NSF 1910284, 1946391, 2147375, and NIH P20GM139768.

Contact Email: akomandu@uark.edu