Causal Diffusion Autoencoders: Toward Representation-Enabled Counterfactual Generation via Diffusion Probabilistic Models

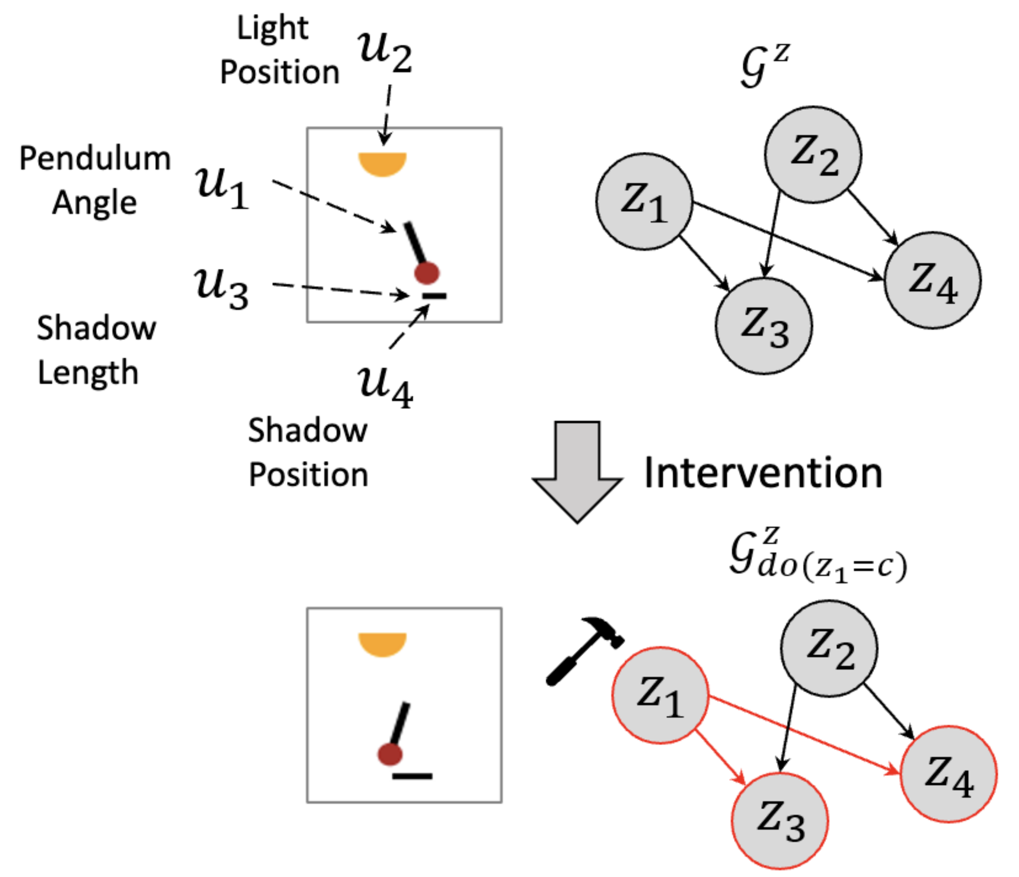
Aneesh Komanduri[1], Chen Zhao[2], Feng Chen[3], and Xintao Wu[1]

[1]University of Arkansas, [2]Baylor University, [3]University of Texas at Dallas

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation

- Diffusion models have become state-of-the-art in image generation, but do not model complex causal dependencies in latent space
- Representation-enabled **counterfactual generation** [1] is an underexplored area in diffusion models
- **Applications:** Healthcare, medicine, biology, etc.



Contributions

- We propose **CausalDiffAE**, a causal representation learning diffusion-based framework to achieve counterfactual generation. We
  - learn a causal representation via stochastic encoder and model causal mechanisms via neural networks in the latent space
  - formulate a variational objective with an alignment prior to ensure disentangled representations
  - propose DDIM for counterfactual generation subject to interventions
  - explore a weak supervision scenario with limited label supervision to enable granular control over generated counterfactuals

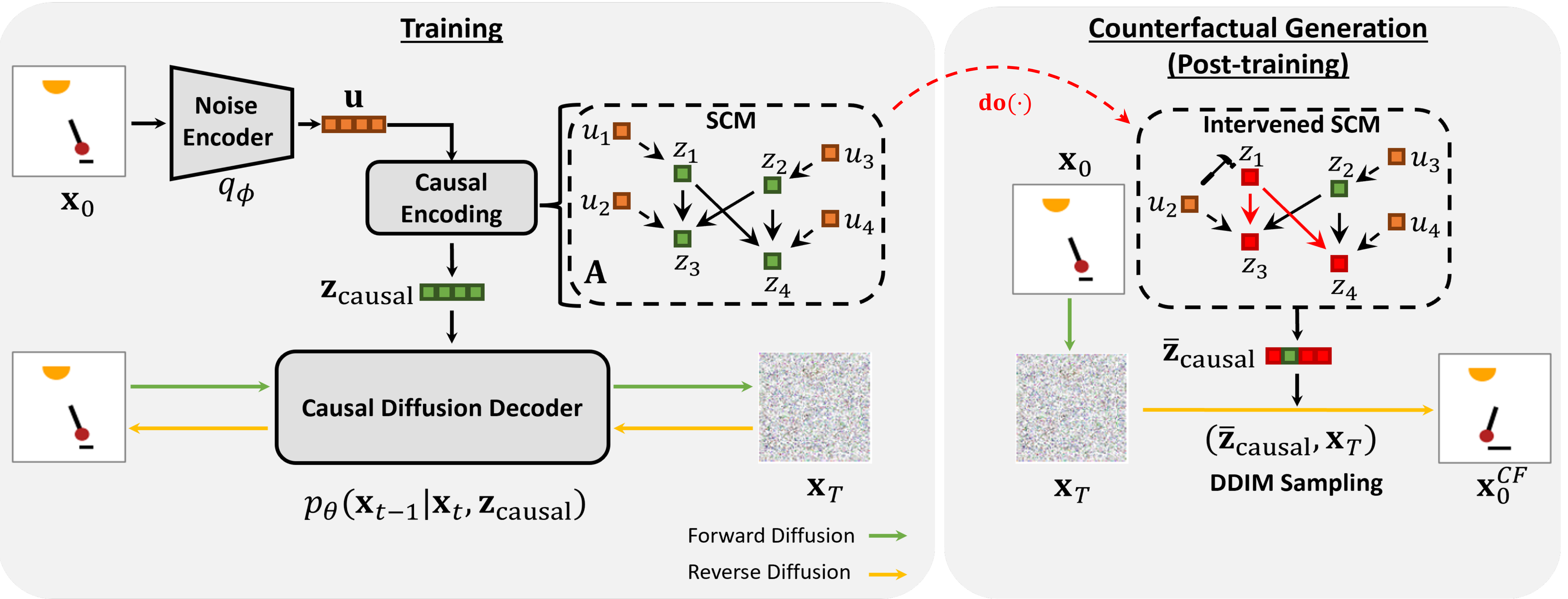
Acknowledgements

This work is supported in part by NSF 1910284, 1946391, 2147375, and NIH P20GM139768

References

[1] A. Komanduri et al. From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling. TMLR. 2024.

Causal Diffusion Autoencoder



Loss

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

Prior

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

Causal Model

$$z_i = f_i(z_{\mathbf{pa}_i}, u_i)$$
$$\mathbf{z} = (\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{u}$$
$$z_i = f_i(\mathbf{A}_i \odot \mathbf{z}; \nu_i) + u_i$$

Counterfactual DDIM

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

Counterfactual Generation Algorithm

Input: Factual sample $\mathbf{x}_0$, intervention target set $\mathcal{I}$ with intervention values c, noise predictor ϵ_θ, encoder ϕ

Output: Counterfactual sample $\mathbf{x}_0^{CF}$

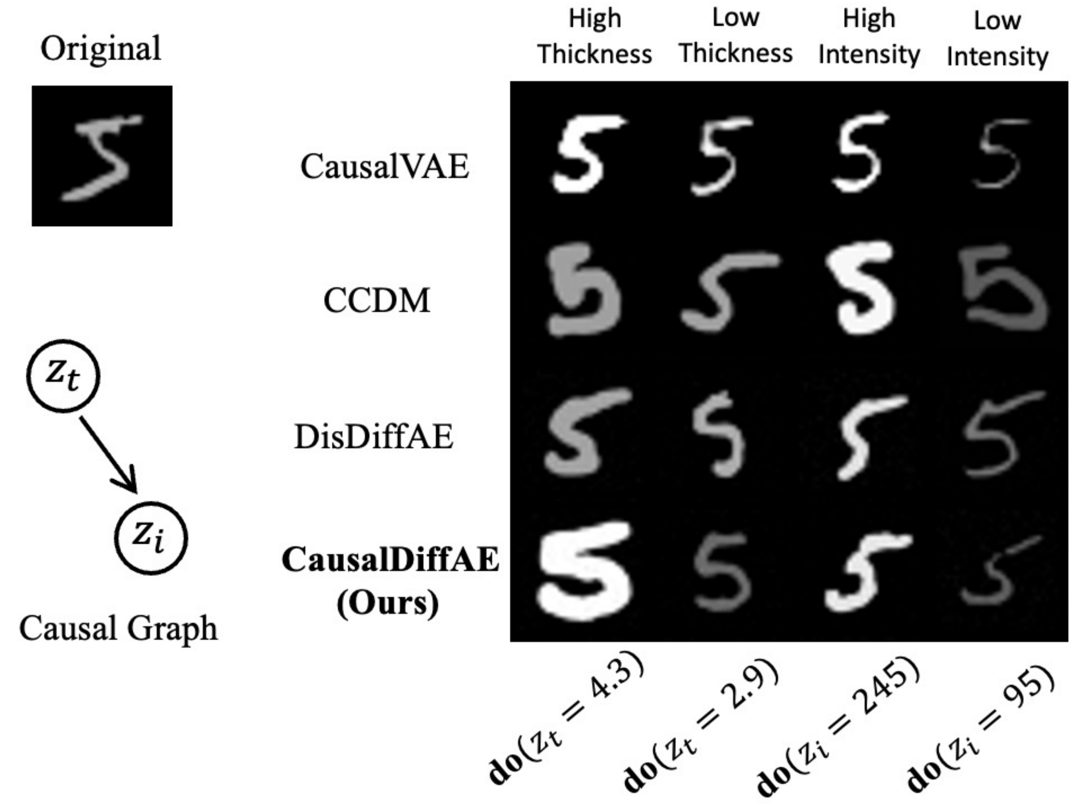
```
u ~ q_φ(u|x_0)                                   ▷ Noise encoding
for i = 1 to n do                                ▷ in topological order
    if i ∈ I then
        z_i = c_i
    else
        z_i = f_i(u_i, z_pa_i)
    end if
end for
z̄_causal = {z_1, ..., z_n}                      ▷ Intervened representation
x_T ~ N(√(α_T) x_0, (1 − α_T) I)
x_T^CF = x_T
for t = T, ..., 1 do                             ▷ DDIM sampling
    x_{t-1}^CF = √(α_{t-1}) ( (x_t^CF − √(1−α_t) ε_θ(x_t^CF, t, z_causal)) / √(α_t) )
                 + √(1 − α_{t-1}) ε_θ(x_t^CF, t, z_causal)
end for
return x_0^CF
```

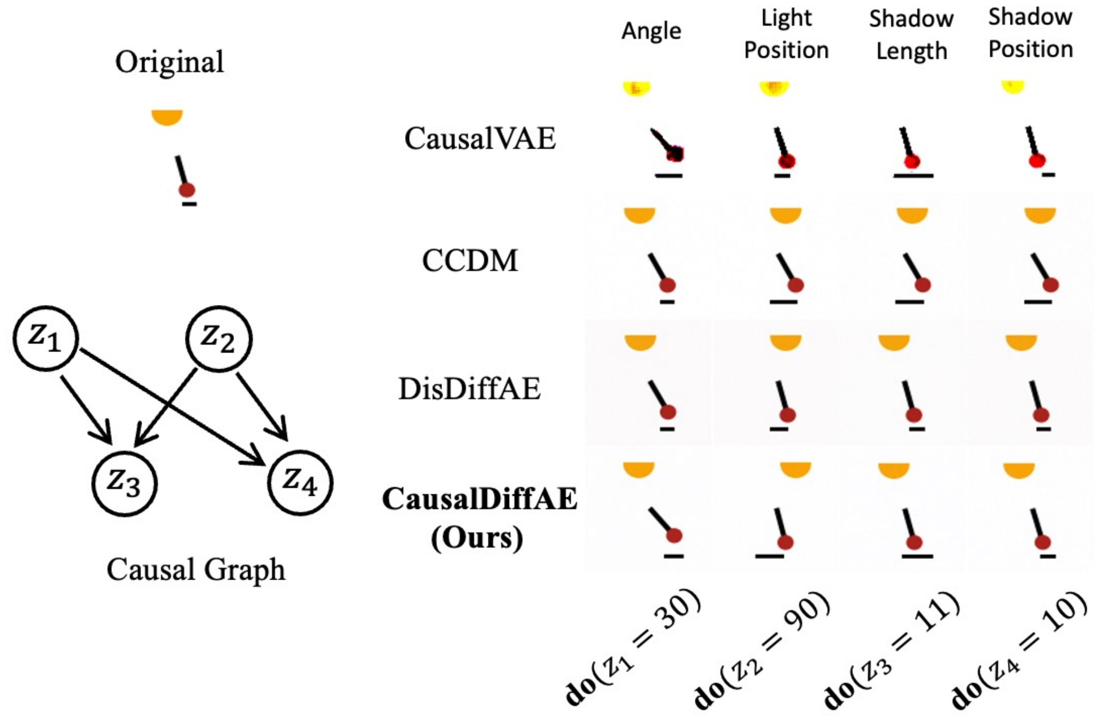
General Procedure:

1. Noise abduction from factual
2. Intervene on causal latents
3. Propagate effects
4. Decode using DDIM

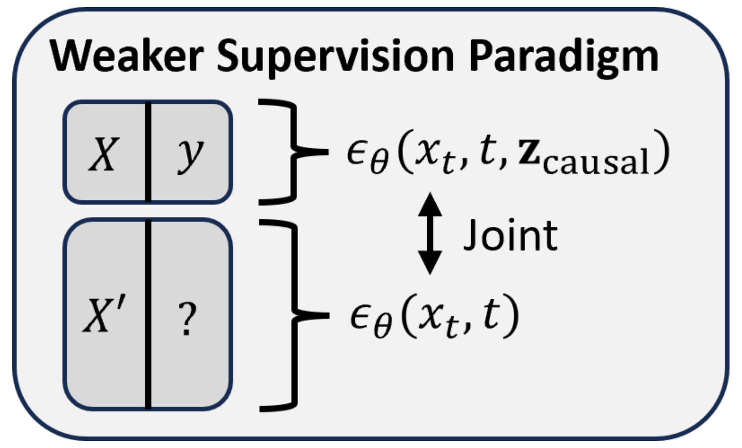
Experimental Evaluation

- **Datasets:** MorphoMNIST and Pendulum, **Metrics:** DCI, Effectiveness, Visual Inspection
- **Generated Counterfactuals**





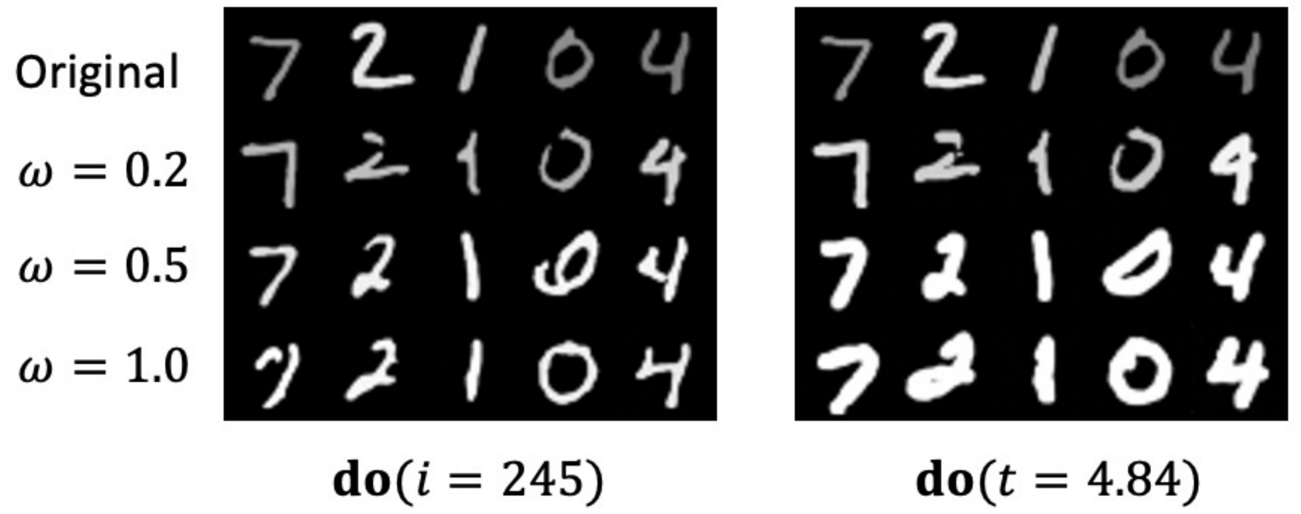
- **Weaker Supervision Training**



Limited label scenario

→ Jointly train conditional & unconditional model

Granular Counterfactual Control



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

- **Quantitative Metrics**

High Disentanglement

Dataset	Model	DCI ↑
MorphoMNIST	CausalVAE	$0.7838_{0.01}$
	DiffAE	$0.3578_{0.01}$
	CausalDiffAE (Ours)	$\mathbf{0.9934_{0.01}}$
Pendulum	CausalVAE	$0.8850_{0.01}$
	DiffAE	$0.3525_{0.01}$
	CausalDiffAE (Ours)	$\mathbf{0.9995_{0.01}}$

High Effectiveness

MorphoMNIST	Thickness MAE ↓		Intensity MAE ↓	
Model	$\mathbf{do}(t)$	$\mathbf{do}(i)$	$\mathbf{do}(t)$	$\mathbf{do}(i)$
CausalVAE	$3.763_{0.01}$	$4.645_{0.01}$	$13.233_{0.01}$	$15.087_{0.01}$
DisDiffAE	$\mathbf{0.377_{0.02}}$	$0.326_{0.02}$	$0.794_{0.02}$	$0.262_{0.02}$
CausalDiffAE	$0.392_{0.02}$	$\mathbf{0.309_{0.02}}$	$\mathbf{0.503_{0.01}}$	$\mathbf{0.256_{0.01}}$

Pendulum	Angle MAE ↓			
Model	$\mathbf{do}(a)$	$\mathbf{do}(l)$	$\mathbf{do}(sl)$	$\mathbf{do}(sp)$
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**

*for more results, see full paper