Causal Manifold Transport for Identifiable Causal Generation in Diffusion Models

Junghyo Sohn[1] , **Wootaek Jeong**[1] , **Sujeong Song**[1] , **Jee Seok Yoon**[2] and **Heung-Il Suk**[1*]
[1] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]SK hynix
{jhsohn0633, wtjeong, sujeongsong, wltjr1007, hisuk}@korea.ac.kr

Abstract

Identifying meaningful latent representations within diffusion models remains a challenging problem for causal approaches. We propose *Causal Manifold Transport Diffusion Model (CMT-Diff)*, a framework that operationalizes causal actions as geometric transformations. By adopting the perspective of backtracking counterfactuals, we formulate the generative process as a composite diffeomorphism that couples the Probability Flow ODE with a Continuous Normalizing Flow. This mapping constructs an exogenous manifold where causal factors align with coordinate variations. Within this geometry, we derive *Causal Manifold Transport (CMT)* to realize interventions as linear vector translations along factor-aligned directions. We establish theoretical identifiability guarantees and demonstrate that our approach facilitates controllable generation by capturing the underlying causal manifold.

1 Introduction

Diffusion models have emerged as the dominant paradigm in modern generative modeling [Song *et al.*, 2021; Yang *et al.*, 2023]. Despite their success in image generation [Yang *et al.*, 2023; Croitoru *et al.*, 2023] and editing [Huang *et al.*, 2024; Shuai *et al.*, 2024], interpreting them through a causal lens remains an open challenge. The primary obstacle is the unstructured nature of their latent spaces. While some studies identify semantic directions [Park *et al.*, 2023; Kwon *et al.*, 2023; Hahm *et al.*, 2024], these do not constitute causal interventions. Unlike Structural Causal Models (SCMs) defined by explicit variables and equations [Pearl, 2009], diffusion latents are high-dimensional and entangled. Consequently, standard diffusion models lack explicit causal variables, rendering intervention operations ill-defined.

While recent works integrate causality into diffusion models [Sanchez and Tsaftaris, 2022; Komanduri *et al.*, 2024; Chao *et al.*, 2024], they primarily rely on external priors or guidance classifiers. Consequently, causal action is not defined and learned within the diffusion model's own latent geometry, limiting mechanism-consistent control. To address this, we posit that a rigorous definition of *causal action* is required within the latent space itself. However, defining such an action is non-trivial and necessitates a formal methodology to operationalize interventions on the learned manifold.

To operationalize interventions on the causal manifold, we distinguish between intervention paradigms. While structural interventions necessitate modifying the mechanism itself, *backtracking counterfactuals* [Von Kügelgen *et al.*, 2023] attribute counterfactual outcomes to variations in exogenous variables under an invariant mechanism. We adopt this formulation as it theoretically validates performing interventions within a fixed generative model. By framing counterfactuals as changes in exogenous inputs rather than structural rules, this perspective justifies treating causal actions directly as latent variable manipulations.

In this work, we propose *CMT-Diff (Causal Manifold Transport Diffusion Model)*, a framework that operationalizes causal interventions as geometric transport operations. To establish a structured causal geometry in the unstructured latent space, we couple the Probability Flow ODE (PF-ODE) [Song *et al.*, 2021] with a Continuous Normalizing Flow (CNF) [Grathwohl *et al.*, 2019]. This invertible mapping serves to reparameterize the unstructured diffusion latent space into a structured exogenous manifold, explicitly modeled to align coordinate variations with distinct causal factors. Within this geometry, we introduce *Causal Manifold Transport (CMT)*. By leveraging intervention-paired observations, we identify factor-specific vector directions in the exogenous space and realize interventions as linear translations. This formulation enables intuitive and identifiable control over the generative process by manipulating latent coordinates along learned causal directions.

Contributions. We formulate a geometric framework that embeds causal structure directly into the diffusion latent space. By coupling the PF-ODE with a CNF, we construct an invertible reparameterization that induces an exogenous coordinate system explicitly aligned with causal variations. Based on this construction, we derive *CMT*, a methodology that operationalizes abstract causal interventions as precise vector translations within the exogenous manifold. We establish theoretical guarantees grounded in latent causal model theory, proving that the learned exogenous representation is identifiable up to diffeomorphism under atomic interventions. Finally, we empirically validate that our geometric formulation accurately captures the underlying causal topology, enabling controlled generation.

*Corresponding author

2 Background

Structural Causal Model (SCM). Pearl's Causal Hierarchy [Pearl, 2009] categorizes causal reasoning into three levels (i.e., observation, intervention, and counterfactuals) offering a systematic method to analyze variable dependencies, identify causal effects, and infer hypothetical scenarios. Building on this hierarchical framework, SCM [Peters *et al.*, 2017] was introduced as a formal method to represent and analyze causal relationships among variables. An SCM $\mathfrak{C} = (S, P_{\boldsymbol{\epsilon}})$ over a set $\mathbf{z}$ of d random variables $z_1, ..., z_d$ consists of (1) a set S of assignments (the structural equations), $S = \{z_i := f_i(PA_i, \epsilon_i)\}$, $i = 1, ..., d$ where f_i is a deterministic function computing each variable z_i from its causal parents $PA_i \subseteq \{z_1, ..., z_d\} \backslash \{z_i\}$ and an exogenous noise variable ϵ_i; (2) a joint distribution $P_{\boldsymbol{\epsilon}}(\epsilon_1, ..., \epsilon_d)$ over the exogenous noise variables, which we require to be jointly independent. In the context of SCMs, two key assumptions underpin our work:

Assumption 1 (Acyclicity)**.** *The causal structure represented by the induced graph G forms a Directed Acyclic Graph (DAG), preventing circular causality.*

Assumption 2 (Causal Sufficiency)**.** *There are no hidden confounders in the model, ensuring the joint independence of the exogenous variables ϵ_i: $P_{\boldsymbol{\epsilon}}(\epsilon_1, \ldots, \epsilon_d) = P_{\boldsymbol{\epsilon}_1}(\epsilon_1) \times \cdots \times P_{\boldsymbol{\epsilon}_d}(\epsilon_d)$.*

Assumptions 1 and 2 establish a unique mapping from exogenous variables $\boldsymbol{\epsilon}$ to endogenous variables $\mathbf{z}$ through a reduced-form function $f(\cdot)$. The induced distribution $P_{\mathbf{z}}$ is defined as the pushforward of $P_{\boldsymbol{\epsilon}}$ through $f(\cdot)$; that is, if $\boldsymbol{\epsilon} \sim P_{\boldsymbol{\epsilon}}$, then $\mathbf{z} = f(\boldsymbol{\epsilon}) \sim P_{\mathbf{z}}$. See Appendix for details of the reduced-form function characteristic.

Backtracking Counterfactuals. Unlike standard interventional counterfactuals that modify the structural mechanism via the do-operator (i.e., $f \to f^{\text{do}}$), *backtracking counterfactuals* preserve the generative mechanism f as invariant. Instead, they attribute counterfactual outcomes to perturbations in the exogenous variables ($\boldsymbol{\epsilon}^{\text{obs}} \to \tilde{\boldsymbol{\epsilon}}$). The counterfactual is thus realized by propagating the modified noise through the fixed model:

$$\tilde{\mathbf{z}} := f(\tilde{\boldsymbol{\epsilon}}). \tag{1}$$

This formulation ensures that interventions respect the learned regularities of the pre-trained model.

3 Problem Formulation

Setup. Let $\mathcal{X}$ denote the observation space and $\mathcal{Z}$ be the latent space associated with a diffusion model. We employ the PF-ODE as a *deterministic* generator $g : \mathcal{Z} \to \mathcal{X}$ and keep its parameters fixed throughout. To endow the unstructured diffusion latent space with an *exogenous* coordinate system, we introduce a learnable, invertible reparameterization $f : \mathcal{E} \to \mathcal{Z}$ (implemented as a CNF). We thus define the composite generative map as:

$$\mathbf{x} = h(\boldsymbol{\epsilon}) := (g \circ f)(\boldsymbol{\epsilon}), \qquad \boldsymbol{\epsilon} \in \mathcal{E},\ \mathbf{x} \in \mathcal{X}. \tag{2}$$

Intuitively, $\mathcal{E}$ serves as an exogenous chart where abstract causal factors are aligned with *simple coordinate movements*, allowing for geometric manipulation.

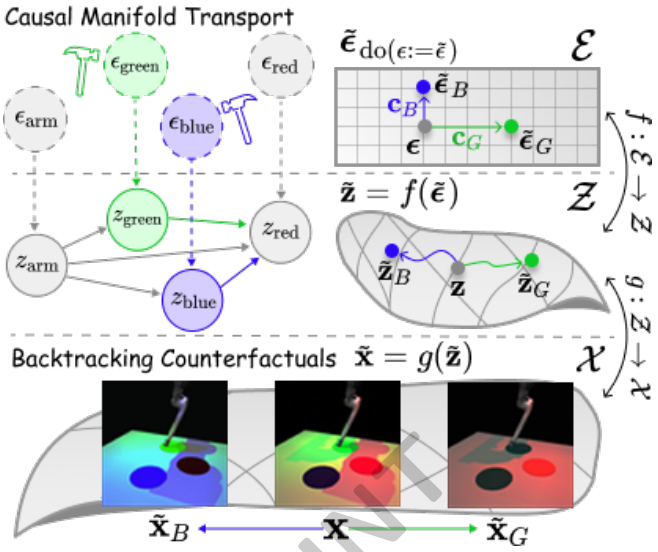

Figure 1: **Conceptual Illustration of Causal Manifold Transport.** The schematic depicts the mechanism of operationalizing causal actions. (Top) In the exogenous manifold $\mathcal{E}$, causal factors are manipulated via linear vector translations along a causal basis. (Middle) These translations are propagated through the endogenous space $\mathcal{Z}$, preserving the causal structure. (Bottom) The transported vectors are decoded to synthesize backtracking counterfactuals with target-specific attribute changes.

Example (Backtracking Counterfactual). We assume access to a dataset of intervention-paired observations $\mathcal{D} = \{(\mathbf{x}, \bar{\mathbf{x}}, \mathcal{I})\}$, where $\mathbf{x}$ is a factual sample, $\bar{\mathbf{x}}$ is the corresponding post-intervention outcome, and $\mathcal{I} \in \{0, 1\}^N$ is an indicator vector denoting which of the N factors is intervened (in the atomic setting). Given a factual observation $\mathbf{x}$, we first infer its endogenous and exogenous coordinates via the inverse mappings:

$$\mathbf{z} = g^{-1}(\mathbf{x}), \qquad \boldsymbol{\epsilon} = f^{-1}(\mathbf{z}). \tag{3}$$

As illustrated in Figure 1, a backtracking counterfactual is produced by applying a parameterized transport operator $\delta_\eta(\cdot)$ to the exogenous code $\boldsymbol{\epsilon}$ and decoding the result through the composite mechanism h:

$$\tilde{\boldsymbol{\epsilon}} = \delta_\eta(\boldsymbol{\epsilon}, \mathcal{I}, \gamma) \ := \ \boldsymbol{\epsilon} + \gamma\, \mathbf{c}_i, \qquad \tilde{\mathbf{x}} = h(\tilde{\boldsymbol{\epsilon}}), \tag{4}$$

where $i = \arg\max_k \mathcal{I}_k$ identifies the target factor, $\gamma \in \mathbb{R}$ controls the intervention magnitude, and η denotes the learnable parameters defining the causal basis $\{\mathbf{c}_k\}_{k=1}^N$.

Goal. Our objective is to jointly learn (i) the invertible exogenous chart $f(\cdot)$ and (ii) the set of factor-aligned directions $\{\mathbf{c}_i\}_{i=1}^N$ in $\mathcal{E}$. The learning criterion is that for each paired sample $(\mathbf{x}, \bar{\mathbf{x}}, \mathcal{I})$, the transported reconstruction $\tilde{\mathbf{x}}$ must accurately match the ground-truth counterfactual $\bar{\mathbf{x}}$ while preserving non-target attributes. The remainder of this paper details the construction of these directions *intrinsic* to the diffusion framework and the training procedure using intervention-paired data.

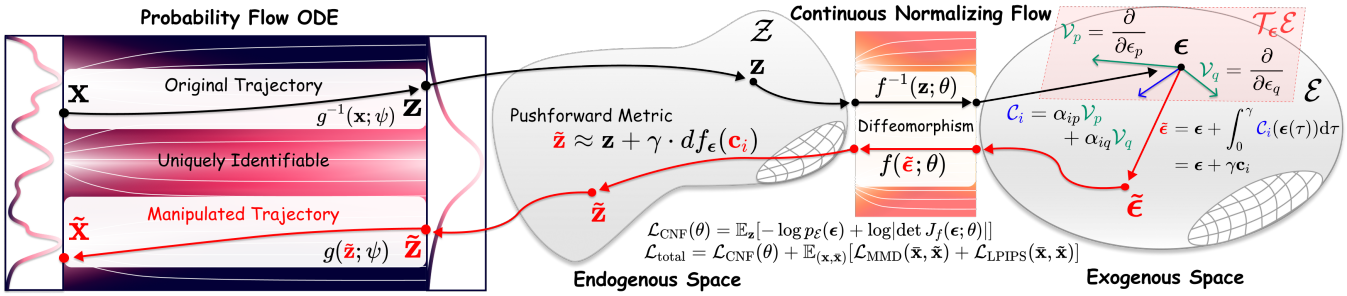

Figure 2: **Overview of CMT-Diff.** The generator is the composite map $h = g \circ f$, where $g : \mathcal{Z} \to \mathcal{X}$ is the PF-ODE and $f : \mathcal{E} \to \mathcal{Z}$ is the CNF. The intervention is formalized as an exogenous translation $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \gamma \mathbf{c}_i$ along a factor-aligned direction $\mathbf{c}_i$ in $\mathcal{E}$. The transport direction is propagated to $\mathcal{Z}$ via the pushforward $df_{\boldsymbol{\epsilon}}$, and the intervened latent $\tilde{\mathbf{z}}$ is decoded by g_ψ to synthesize the backtracking counterfactual $\tilde{\mathbf{x}}$.

4 Causal Manifold Transport Diffusion Models

4.1 Elucidating the Diffusion via a Causal Lens

Endogenous and Exogenous Spaces. In this study, we define the latent space of the PF-ODE $g : \mathcal{Z} \to \mathcal{X}$ as the endogenous space $\mathcal{Z}$. Although $\mathcal{Z}$ encodes causal information, the latent space in diffusion models is not inherently disentangled, making it challenging to identify the directional relationships between causal factors. To address this and extract independent exogenous variables, we introduce an invertible CNF $f : \mathcal{E} \to \mathcal{Z}$ that reparameterizes the diffusion latent space. Its inverse $f^{-1} : \mathcal{Z} \to \mathcal{E}$ maps an endogenous latent $\mathbf{z}$ to exogenous coordinates $\boldsymbol{\epsilon}$, i.e., $\boldsymbol{\epsilon} = f^{-1}(\mathbf{z})$. By constructing the CNF as a diffeomorphic map [Grathwohl *et al.*, 2019], $\mathcal{Z}$ and $\mathcal{E}$ share the same topological structure.

Smooth Manifold and Diffeomorphism. To formalize the properties of $\mathcal{Z}$ and $\mathcal{E}$, we present the following Observation 1 and Theorem 1 from [Dominguez-Olmedo *et al.*, 2023]:

Observation 1. *Under causal sufficiency, if each marginal distribution $P_{\boldsymbol{\epsilon}_i}$ has a d_i-dimensional smooth manifold support, the exogenous manifold $\mathcal{E}$ forms a smooth manifold of dimension $d = \sum_i d_i$.*

Theorem 1. *For an acyclic SCM with the exogenous manifold $\mathcal{E}$ and the endogenous manifold $\mathcal{Z}$, if the following conditions hold: (1) $\mathcal{E}$ is a smooth manifold; (2) $f_i(\cdot)$ is differentiable with $\partial_{\epsilon_i} f_i(\epsilon_i) \neq 0$ and is injective with respect to ϵ_i; then $\mathcal{Z}$ is a smooth manifold and the mapping $f : \mathcal{E} \to \mathcal{Z}$ is a diffeomorphism.*

Under Assumption 2, $\mathcal{E}$ is a smooth manifold, characterized by its locally Euclidean structure and differentiable properties (Observation 1). The CNF $f(\cdot)$, satisfying the second condition of Theorem 1, ensures that both $\mathcal{Z}$ and $\mathcal{E}$ are smooth manifolds and that $f : \mathcal{E} \to \mathcal{Z}$ is a diffeomorphism. This diffeomorphism preserves the topological and differential structures between $\mathcal{Z}$ and $\mathcal{E}$, thereby conserving local geometric properties under the mapping $f_\theta(\cdot)$. Thus, geometric transports in $\mathcal{E}$ induce well-defined, causally consistent transformations in $\mathcal{Z}$.

4.2 The Causal Manifold Transport

Defining Transformations via Vector Fields. A vector field $\mathcal{V}$ on a manifold $\mathcal{E}$ is a smooth mapping assigning each point $\boldsymbol{\epsilon} \in \mathcal{E}$ a tangent vector $\mathcal{V}(\boldsymbol{\epsilon}) \in \mathcal{T}_{\boldsymbol{\epsilon}}\mathcal{E}$, representing the directional adjustment at that point. In our context, we aim to define vector fields that encode interventions on specific causal factors. We first invoke Theorem 2 (adapted from [Lee, 2012]) to establish the existence of a coordinate system aligned with such fields:

Theorem 2. *Let $\mathcal{E}$ be a d-dimensional smooth manifold. Consider a set of linearly independent vector fields $\{\mathcal{V}_1, \ldots, \mathcal{V}_d\}$ on an open subset $\mathcal{U} \subset \mathcal{E}$ that commute (i.e., their Lie bracket vanishes). Then, for any $\boldsymbol{\epsilon} \in \mathcal{U}$, there exists a smooth coordinate chart centered at $\boldsymbol{\epsilon}$ such that $\frac{\partial}{\partial \epsilon_i} = \mathcal{V}_i$ for all $i = 1, \ldots, d$.*

While Observation 1 suggests that the exogenous manifold $\mathcal{E}$ admits local Euclidean approximations, our architecture enforces a stronger geometric condition. By construction, the CNF establishes a global diffeomorphism between $\mathcal{E}$ and the latent space $\mathbb{R}^d$. This topological equivalence allows us to utilize a global coordinate system where the standard basis $\{\mathbf{e}_j\}_{j=1}^d$ constitutes a valid frame for the tangent space everywhere. Consequently, a geometric transport along a basis vector creates a globally consistent linear trajectory, formalized as:

$$\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \int_0^\gamma \mathcal{V}_j(\boldsymbol{\epsilon}(\tau))\, \mathrm{d}\tau = \boldsymbol{\epsilon} + \gamma \mathbf{e}_j, \tag{5}$$

where γ denotes the transport magnitude.

Establishing a Causal Basis for Transport. The standard basis transformation in Eq. (5) manipulates latent dimensions but does not necessarily align with underlying causal factors. To address this, we postulate that N causal factors are encoded within $\boldsymbol{\epsilon} \in \mathbb{R}^d$ ($N \leq d$) and construct a dedicated *Causal Vector Field*. Inspired by distributed representations [Hinton *et al.*, 1984], we employ Proposition 1 to define this field:

Proposition 1 (Causal Vector Fields)**.** *Based on Theorem 2, we define a set of new vector fields $\{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ on $\mathcal{E}$. Each causal vector field $\mathcal{C}_i$ is constructed as a linear combination of the standard coordinate fields $\{\mathcal{V}_j\}_{j=1}^d$:*

$$\mathcal{C}_i = \sum_{j=1}^{d} \alpha_{ij} \mathcal{V}_j, \quad \text{for } i = 1, \ldots, M, \tag{6}$$

where $\alpha_{ij} \in \mathbb{R}$ are learnable scalar coefficients. If linearly independent, these fields span an M-dimensional causal subspace of $\mathcal{T}_{\mathcal{E}}$.

Corollary 1 (Causal Basis)**.** *The constant vector fields $\mathcal{C}_i$ defined in Proposition 1 correspond uniquely to constant vectors $\mathbf{c}_i \in \mathbb{R}^d$, defined as $\mathbf{c}_i = \sum_{j=1}^{d} \alpha_{ij}\mathbf{e}_j$. If $N = M$, the set $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ forms the* Causal Basis *for the latent space.*

By establishing these fields, we define *CMT* as the integration along the causal vector field $\mathcal{C}_i$. Specifically, for a target factor i, the intervention is formalized as:

$$\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \int_0^{\gamma} \mathcal{C}_i(\boldsymbol{\epsilon}(\tau))\,\mathrm{d}\tau = \boldsymbol{\epsilon} + \gamma \mathbf{c}_i. \tag{7}$$

While Eq. (7) simplifies to a linear translation due to the global Euclidean structure of $\mathcal{E}$, its validity is rooted in the factor-aligned direction $\mathbf{c}_i$. The CNF maps the entangled endogenous space $\mathcal{Z}$ to this disentangled manifold $\mathcal{E}$, ensuring that linear transports along $\mathbf{c}_i$ correspond to meaningful, mechanism-preserving interventions. See Appendix for proofs of Proposition 1 and Corollary 1.

Pushforward Operation. CMT is characterized by integrating a causal vector field on the exogenous manifold $\mathcal{E}$. We investigate how this vector field propagates through the composite generative mapping $h = g \circ f$, where $f : \mathcal{E} \to \mathcal{Z}$ denotes the CNF and $g : \mathcal{Z} \to \mathcal{X}$ is the PF-ODE. For a given point $\boldsymbol{\epsilon} \in \mathcal{E}$, a causal basis vector $\mathbf{c} \in \mathcal{T}_{\boldsymbol{\epsilon}}\mathcal{E}$—derived from $\{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$—is first mapped to the tangent space $\mathcal{T}_{f(\boldsymbol{\epsilon})}\mathcal{Z}$ via the pushforward operator $df_{\boldsymbol{\epsilon}}$. Here, $df_{\boldsymbol{\epsilon}}(\mathbf{c})$ represents the induced transport direction in the endogenous space $\mathcal{Z}$. Letting $\mathbf{z} = f(\boldsymbol{\epsilon})$, this vector is subsequently pushed forward to the tangent space $\mathcal{T}_{g(\mathbf{z})}\mathcal{X}$ of the observation space $\mathcal{X}$ via $dg_{\mathbf{z}}$. Consequently, the term $dg_{\mathbf{z}}(df_{\boldsymbol{\epsilon}}(\mathbf{c}))$ captures the infinitesimal effect of the initial causal basis vector $\mathbf{c}$ after successive transformations through the CNF and PF-ODE. By the chain rule, we have

$$d(g \circ f)_{\boldsymbol{\epsilon}}(\mathbf{c}) = dg_{f(\boldsymbol{\epsilon})}(df_{\boldsymbol{\epsilon}}(\mathbf{c}))\,, \tag{8}$$

which explicitly shows how the transport vector field is pushed forward from $\mathcal{E}$ to $\mathcal{Z}$, and ultimately to $\mathcal{X}$. Figure 2 illustrates the overall pipeline of our causality-integrated diffusion framework, *CMT-Diff*.

4.3 Identifiability Analysis

We ground our theoretical validity in the Latent Causal Model (LCM) framework [Brehmer *et al.*, 2022]. We first recall the standard definition of LCMs and their identifiability conditions under hard interventions, and then extend this guarantee to our soft intervention setting.

Definition 1 (Latent Causal Models)**.** *An LCM is a tuple $\mathcal{Q} = (\mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}})$, comprising an acyclic SCM $\mathcal{C}$ over latent variables $\mathcal{Z}$, an observation space $\mathcal{X}$, a diffeomorphism $g : \mathcal{Z} \to \mathcal{X}$, and a set of interventions $\mathcal{I}$ with distribution $p_{\mathcal{I}}$. Two LCMs $\mathcal{Q}, \mathcal{Q}'$ are* equivalent *($\mathcal{Q} \sim \mathcal{Q}'$) if there exists an isomorphism preserving the causal structure and interventional distributions.*

Theorem 3 (Identifiability of Real-Valued LCMs [Brehmer *et al.*, 2022])**.** *For LCMs $\mathcal{Q}, \mathcal{Q}'$ with real-valued variables and atomic hard interventions, equivalence $\mathcal{Q} \sim \mathcal{Q}'$ holds if and only if their weakly supervised distributions are identical: $p^{\mathcal{X}}_{\mathcal{Q}}(\mathbf{x}, \tilde{\mathbf{x}}) = p^{\mathcal{X}}_{\mathcal{Q}'}(\mathbf{x}, \tilde{\mathbf{x}})$.*

Identifiability under Causal Manifold Transport. Our framework instantiates an LCM where the mapping g is the PF-ODE and the latent structure is governed by a CNF $f : \mathcal{E} \to \mathcal{Z}$. While Theorem 3 assumes hard interventions, our Causal Manifold Transport operates as a soft intervention on exogenous variables $\boldsymbol{\epsilon}$. We demonstrate that the identifiability guarantee extends to this relaxation under the intervention-paired setting. Specifically, we formalize the conditions under which soft geometric shifts enable the recovery of the true causal topology:

Proposition 2 (Identifiability under Atomic Soft Exogenous Shifts)**.** *Let $\mathcal{Q}, \mathcal{Q}'$ be LCMs with diffeomorphic observation maps. Assume intervention-paired data $(\mathbf{x}, \tilde{\mathbf{x}}, \mathcal{I})$ generated by atomic soft shifts $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \gamma \mathbf{c}_i$. If $\mathcal{Q}$ and $\mathcal{Q}'$ induce identical distributions $p(\mathbf{x}, \tilde{\mathbf{x}}, \mathcal{I})$, then their latent representations are equivalent up to a diffeomorphism that preserves the intervention structure.*

Proposition 3 (Counterfactual Uniqueness via Identifiable Transport)**.** *Consider two equivalent LCMs $\mathcal{Q}$ and $\mathcal{Q}'$ that are identifiable up to an intervention-preserving diffeomorphism (as established in Proposition 2). Under the strict determinism and bijectivity of the generative mechanism $h = g \circ f$, equivalence in the latent causal representation implies equivalence in counterfactual generation. Specifically, for any observation $\mathbf{x}$ and intervention target $\mathcal{I}$, the generated counterfactual is invariant to the specific choice of the identified model: $\tilde{\mathbf{x}}_{\mathcal{Q}} = \tilde{\mathbf{x}}_{\mathcal{Q}'}$.*

Collectively, Propositions 2 and 3 establish that *under the assumptions of deterministic and bijective h and intervention-preserving equivalence*, recovering the causal transport topology in $\mathcal{E}$ is sufficient to make the induced counterfactual mapping in $\mathcal{X}$ unique within the identifiable equivalence class. See Appendix for proofs of Propositions 2 and 3.

4.4 Training Procedure

Overall Framework. Initially, we extract the endogenous variable $\mathbf{z}$ from the observation $\mathbf{x}$ by inverting the PF-ODE. Leveraging the deterministic nature of the PF-ODE, we define the extraction dynamics as:

$$\mathrm{d}\mathbf{x} = -ts(\mathbf{x}_t, t; \psi)\mathrm{d}t. \tag{9}$$

The endogenous variable $\mathbf{z} = g^{-1}(\mathbf{x}; \psi)$ is obtained by integrating Eq. (9) forward from $t = 0$ to $t = T$ using a standard discretization scheme over time steps $\{t_0, \ldots, t_n\}$. To ensure consistency, the parameters ψ are frozen during this process. We parameterize the CNF as an ODE $du(\tau)/d\tau = v_\theta(u(\tau), \tau)$, which induces an invertible map $f : \mathcal{E} \to \mathcal{Z}$ by integrating from τ_0 to τ_1 with $u(\tau_0) = \boldsymbol{\epsilon}$ and $u(\tau_1) = \mathbf{z}$, i.e., $\mathbf{z} = f(\boldsymbol{\epsilon}; \theta)$. The inverse map $\boldsymbol{\epsilon} = f^{-1}(\mathbf{z}; \theta)$ is obtained by integrating the same ODE backward from τ_1 to τ_0 with terminal condition $u(\tau_1) = \mathbf{z}$.

The manipulation function $\tilde{\boldsymbol{\epsilon}} = \delta(\boldsymbol{\epsilon}, \mathcal{I}, \gamma; \eta)$ maps an original point $\boldsymbol{\epsilon} \in \mathcal{E}$ and an intervention vector $\mathcal{I} \in \{0, 1\}^N$ to a manipulated point $\tilde{\boldsymbol{\epsilon}} \in \mathcal{E}$. Its parameters η include a learnable logit matrix $\mathbf{A} \in \mathbb{R}^{N \times d}$ (where N is the number of causal factors and d is the feature dimension). For each feature j, the vector $\mathbf{A}[:, j] \in \mathbb{R}^N$ contains the logits for the N causal factors' influence on that feature. Within the function $\delta(\cdot; \eta)$,

the process for deriving and using causal basis vectors begins by computing the matrix $\mathbf{G} \in \mathbb{R}^{N \times d}$. Each column $\mathbf{G}[:, j]$ is an N-dimensional vector (typically one-hot or soft one-hot) obtained from Gumbel-Softmax applied to the corresponding logits $\mathbf{A}[:, j] \quad (\text{for } j = 1, \ldots, d)$:

$$\mathbf{G}[:, j] = \text{Gumbel-Softmax}(\mathbf{A}[:, j]). \tag{10}$$

The i-th row of this matrix, $\mathbf{G}[i, :]$, consequently represents a d-dimensional vector of factor i's influence across all features after the Gumbel-Softmax selection process. The d-dimensional causal basis vector $\mathbf{c}_k \in \mathbb{R}^d$ for an individual causal factor k is then defined as the k-th row of $\mathbf{G}$:

$$\mathbf{c}_k = \mathbf{G}[k, :] \quad (\text{for } k = 1, \ldots, N). \tag{11}$$

This construction instantiates the coefficients in Proposition 1 as $\alpha_{ij} := \mathbf{G}_{ij}$, i.e., $\mathcal{C}_i = \sum_{j=1}^{d} \mathbf{G}_{ij} \mathcal{V}_j$. The intervention vector $\mathcal{I}$ determines which features are associated with specific interventions and whether an intervention is applied. If $\mathcal{I}$ indicates an intervention on a specific factor k (e.g., $\mathcal{I}_k = 1$ and all other $\mathcal{I}_j = 0$), the corresponding causal basis $\mathbf{c}_k$ is scaled by γ (a scalar sampled from a uniform distribution) and added to $\boldsymbol{\epsilon}$ to obtain $\tilde{\boldsymbol{\epsilon}} := \boldsymbol{\epsilon} + \gamma \mathbf{c}_k$. Given an intervened exogenous coordinate $\tilde{\boldsymbol{\epsilon}}$, we obtain the intervened endogenous latent by the forward CNF map,

$$\tilde{\mathbf{z}} = f(\tilde{\boldsymbol{\epsilon}}; \theta). \tag{12}$$

Finally, the backtracking counterfactual $\tilde{\mathbf{x}}$ is obtained by decoding the intervened latent $\tilde{\mathbf{z}}$ via the PF-ODE. This process is mathematically equivalent to solving the dynamics defined in Eq. (9) backward in time from $t = T$ to $t = 0$, subject to the terminal boundary condition $\mathbf{x}_T = \tilde{\mathbf{z}}$.

Training Objective. The counterfactual generation framework is optimized using a combined loss function to ensure accurate data distribution modeling and perceptual fidelity in generated counterfactuals. The primary loss component trains the CNF by maximum likelihood on the endogenous latents $z = g^{-1}(x; \psi)$. Using the change-of-variables formula with base density $p_{\mathcal{E}}(\epsilon) = \mathcal{N}(0, I)$ and $\epsilon = f^{-1}(z; \theta)$, the CNF negative log-likelihood is $\mathcal{L}_{\text{CNF}}(\theta) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}}[-\log p_{\mathcal{E}}(f^{-1}(g^{-1}(\mathbf{x}; \psi); \theta)) - \log |\det J_{f^{-1}} g^{-1}(\mathbf{x}; \psi); \theta)|]$. Equivalently, writing $\epsilon = f^{-1}(z; \theta)$, we have

$$\mathcal{L}_{\text{CNF}}(\theta) = \mathbb{E}_{\mathbf{z}} \left[-\log p_{\mathcal{E}}(\boldsymbol{\epsilon}) + \log |\det J_f(\boldsymbol{\epsilon}; \theta)| \right]. \tag{13}$$

For CNFs, the log-determinant term is computed via the instantaneous change-of-variables formula along the ODE trajectory. For counterfactual image generation, the loss also incorporates Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012] for global distributional properties and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] for perceptual similarity. The overall loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CNF}}(\theta) + \mathbb{E}_{(\mathbf{x}, \bar{\mathbf{x}})} \left[\mathcal{L}_{\text{MMD}}(\bar{\mathbf{x}}, \tilde{\mathbf{x}}) + \mathcal{L}_{\text{LPIPS}}(\bar{\mathbf{x}}, \tilde{\mathbf{x}}) \right], \tag{14}$$

where $\tilde{\mathbf{x}}$ is the generated counterfactual and $\bar{\mathbf{x}}$ is the true post-intervention image. Minimizing $\mathcal{L}_{\text{total}}$ trains the CNF to map the intervention manifold. MMD and LPIPS losses assess statistical and perceptual similarities between the generated counterfactuals and the true post-intervention images, while $\mathcal{L}_{\text{CNF}}(\theta)$ maximizes data likelihood for accurate distribution modeling.

Algorithm 1 Overall Training Procedure

Require: Intervention-paired data $\mathcal{D} = \{(\mathbf{x}, \bar{\mathbf{x}}, \mathcal{I})\}$, PF-ODE g_ψ (score net $s(\cdot; \psi)$), CNF f_θ, basis parameters η

Phase 1 (pretrain PF-ODE).
while not converged **do**
  $\psi \leftarrow \psi - \nabla_\psi \mathcal{L}_{\text{EDM}}$ [Karras *et al.*, 2022]
end while
Phase 2 (train CNF & CMT):
while not converged **do**
  $\mathbf{z} \leftarrow g^{-1}(\mathbf{x}; \psi), \quad \boldsymbol{\epsilon} \leftarrow f^{-1}(\mathbf{z}; \theta), \quad \gamma \sim p(\gamma)$
  $\tilde{\boldsymbol{\epsilon}} \leftarrow \delta(\boldsymbol{\epsilon}, \mathcal{I}, \gamma; \eta)$
  $\tilde{\mathbf{x}} \leftarrow g(f(\tilde{\boldsymbol{\epsilon}}; \theta); \psi)$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{CNF}}(\theta) + \mathcal{L}_{\text{LPIPS}}(\tilde{\mathbf{x}}, \bar{\mathbf{x}}) + \mathcal{L}_{\text{MMD}}(\tilde{\mathbf{x}}, \bar{\mathbf{x}})$
  $(\theta, \eta) \leftarrow (\theta, \eta) - \alpha \nabla_{(\theta, \eta)} \mathcal{L}$
end while

5 Experiments

Setup. We utilize the CausalCircuit dataset [Brehmer *et al.*, 2022], which consists of 120K image pairs before and after interventions, split into 100K/10K/10K for training, validation, and testing. To rigorously benchmark performance, we conduct a comprehensive comparison against a wide range of state-of-the-art causal generative models. Our evaluation covers VAE-based approaches, including β-VAE [Burgess *et al.*, 2018], iVAE [Khemakhem *et al.*, 2020], CausalVAE [Yang *et al.*, 2021], SCM-VAE [Komanduri *et al.*, 2022], ILCM [Brehmer *et al.*, 2022], and ICM-VAE [Komanduri *et al.*, 2023], as well as diffusion-based baselines such as DiffAE [Preechakul *et al.*, 2022] and CausalDiffAE [Komanduri *et al.*, 2024]. We report (i) representation quality using DCI [Eastwood and Williams, 2018] and (ii) generation quality using FID [Heusel *et al.*, 2017], IS [Salimans *et al.*, 2016], and SSIM [Wang *et al.*, 2004]. More details regarding the experimental setup, baselines, and evaluation metrics are provided in the Appendix.

Qualitative Results. Figure 3 demonstrates the geometric realization of backtracking counterfactuals. Translating exogenous coordinates along factor-aligned directions induces consistent, monotonic variations in target attributes (e.g., light intensity, arm position) while preserving orthogonal content. This validates our premise that causal interventions can be operationalized as vector transports within the exogenous manifold. Furthermore, composite interventions (e.g., Green and Blue) exhibit additive behavior, confirming that the learned directions constitute an independent causal basis enabling compositional control.

Quantitative Results. Table 1 evaluates structural identifiability using the DCI metric, which quantifies the extent to which each latent dimension captures a single, distinct generative factor. While standard diffusion baselines (e.g., DiffAE) exhibit substantial latent entanglement, CMT-Diff attains a DCI score of 0.99, indicating strong factor alignment in the learned exogenous coordinates. This suggests that the learned transports are factor-specific and consistent with the intervention targets, supporting the use of linear translations in the exogenous manifold for controllable counterfactual generation. Although CausalDiffAE achieves similar performance,

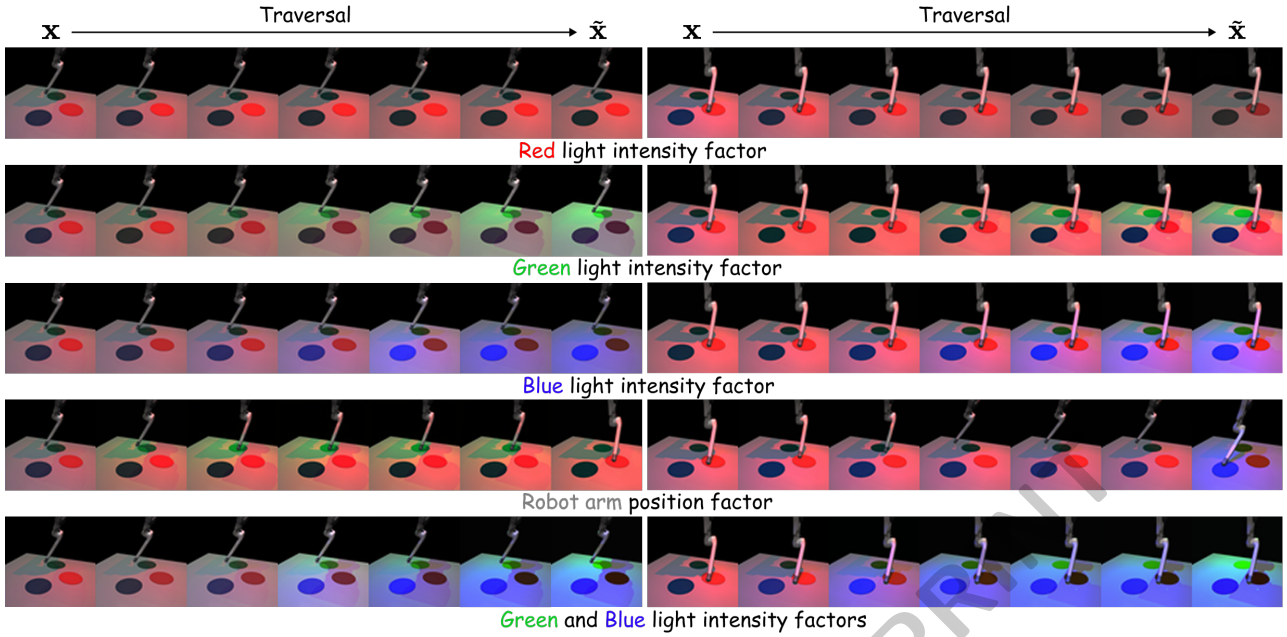

Figure 3: **Qualitative traversals along causal directions.** Each row shows backtracking counterfactual traversals obtained by transporting the inferred exogenous coordinate along a factor-aligned direction.

Method	DCI (↑)
VAE-based Approaches	
β-VAE [Burgess *et al.*, 2018]	0.69
iVAE [Khemakhem *et al.*, 2020]	0.74
CausalVAE [Yang *et al.*, 2021]	0.88
SCM-VAE [Komanduri *et al.*, 2022]	0.86
ILCM [Brehmer *et al.*, 2022]	0.99
ICM-VAE [Komanduri *et al.*, 2023]	0.98
Diffusion-based Approaches	
DiffAE [Preechakul *et al.*, 2022]	0.35
CausalDiffAE [Komanduri *et al.*, 2024]	0.99
CMT-Diff (Ours)	**0.99**

Table 1: **Quantitative comparison of disentanglement.** DCI scores (higher is better) on the CausalCircuit dataset for CMT-Diff and baseline causal generative models.

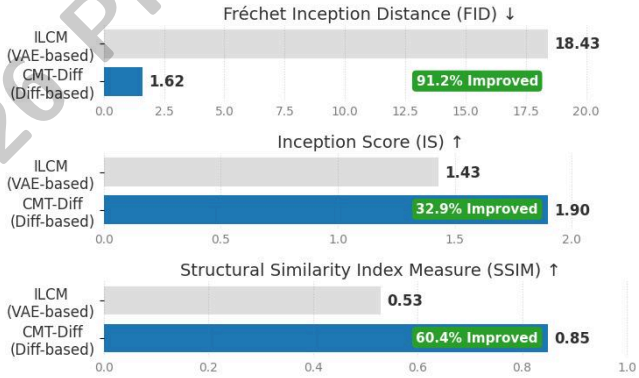

Figure 4: **Quantitative comparison of counterfactual generation quality.** FID, IS, and SSIM on CausalCircuit, comparing counterfactual fidelity between CMT-Diff and ILCM under a comparable intervention-paired setting.

CMT-Diff learns the exogenous manifold intrinsically within the diffusion model using intervention-paired data and labels, whereas CausalDiffAE uses ground-truth exogenous variables as external guidance.

Generative Quality and Model Comparison. A critical premise of *CMT-Diff* is that disentanglement alone is insufficient for high-fidelity causal generation. While VAE-based baselines such as ILCM [Brehmer *et al.*, 2022] achieve competitive factor alignment (Table 1), their reconstruction capabilities often falter in generating realistic counterfactuals. To rigorously evaluate this trade-off, we benchmark against ILCM, which operates under a comparable intervention-paired regime. We explicitly acknowledge that our method leverages auxiliary intervention target labels to define the transport direction. However, this comparison is intended to demonstrate the distinct advantage of integrating target-aware geometric transports within a diffusion backbone. As shown in Figure 4, *CMT-Diff* yields substantial improvements in perceptual fidelity while maintaining precise causal control.

Coherence of Geometric Transport. Figure 6 visualizes the topology of the learned exogenous manifold $\mathcal{E}$ via t-SNE projections. We observe that interventions manifest as coherent, directionally consistent translations applied to local neighborhoods of factual samples, distinctly avoiding arbitrary discontinuities. This empirical regularity corroborates our theoretical premise that causal interventions operate as smooth, factor-specific transport maps within the latent chart.

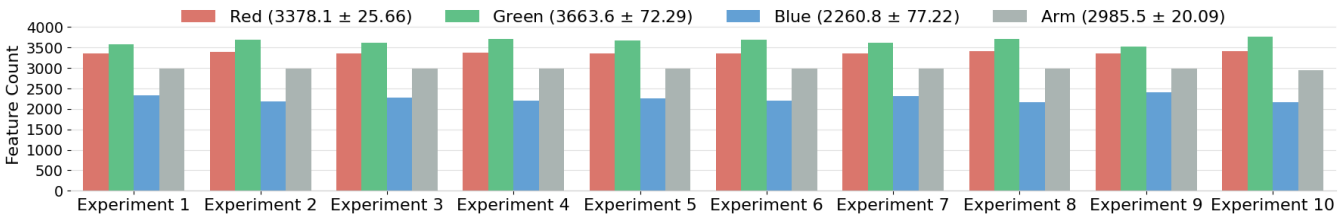

Figure 5: **Causal feature allocation induced by the learned causal basis.** Across 10 independent runs, the number of latent dimensions assigned to each causal factor by the Gumbel-Softmax allocation (mean ± std shown in the legend).

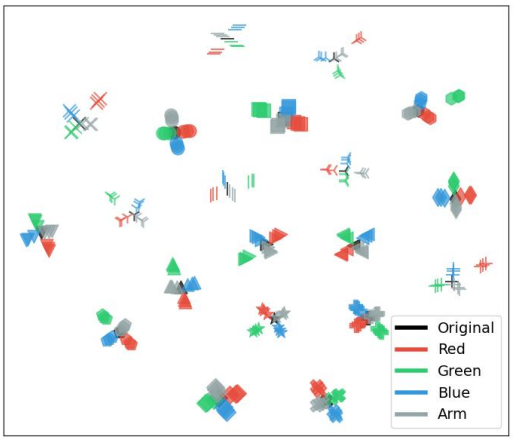

Figure 6: **t-SNE visualization of causal representations.** 2D t-SNE embeddings of inferred exogenous coordinates for factual samples and their intervention-targeted counterparts (Red/Green/Blue/Arm).

Stability of the Causal Basis. Figure 5 probes the robustness of the learned basis construction to random initialization across 10 independent runs. Specifically, we measure how many latent dimensions are assigned to each factor by the Gumbel-Softmax based allocation and report the mean and standard deviation across runs. The allocation exhibits low variation, suggesting that the basis learning procedure is relatively stable and tends to converge to repeatable factor-wise partitions of the exogenous representation in this setting. While this does not by itself establish identifiability, it provides empirical evidence that the learned transport parameterization is not overly sensitive to initialization. In practice, such repeatability is desirable because it reduces the chance that the intervention semantics change drastically across retraining.

6 Related Work

Backtracking Counterfactuals. In SCMs, counterfactual is typically formalized through structural interventions that modify the governing equations (e.g., via the do-operator) [Pearl, 2009]. For trained generative models, however, changing the mechanism defines a different data-generating process than the one represented by the fixed model parameters. Recent theoretical work addresses this mismatch by introducing *backtracking counterfactuals*, which keep the structural equations unchanged and instead explain changes through modifications of exogenous variables [Von Kügelgen *et al.*, 2023]. Follow-up studies refine this idea by adding feasibility or naturalness constraints [Hao *et al.*, 2024] or by framing recourse as an optimization problem [Fatemi *et al.*, 2025]. In contrast to optimization-centric formulations or approaches that require explicit graph constraints, our work operationalizes backtracking inside diffusion models by constructing an explicit exogenous coordinate system and realizing backtracking as geometric transport along factor-aligned directions.

Causal Generative Models. Causal generative modeling has been extensively studied in VAE-based frameworks that learn identifiable latent causal variables using *intervention-paired* observations and related supervision signals (e.g., [Locatello *et al.*, 2020; Brehmer *et al.*, 2022; Ahuja *et al.*, 2022; Lippe *et al.*, 2022; Lippe *et al.*, 2023; Zhang *et al.*, 2023; Zhou *et al.*, 2024]). These works provide identifiability guarantees and practical procedures for recovering causal factors in structured latent spaces. More recently, diffusion models have been explored for causal modeling (e.g., [Sanchez and Tsaftaris, 2022; Komanduri *et al.*, 2024; Chao *et al.*, 2024]), often targeting interventional or counterfactual generation by introducing additional structural information or guidance signals. A key difference is that prior diffusion-based approaches typically impose causality externally to the diffusion latent geometry, for example by assuming explicit structural priors (such as known graphs or mechanism constraints) or by constructing a causal representation with an auxiliary encoder and then conditioning or guiding a diffusion generator. In contrast, we focus on learning an intrinsic causal representation within the diffusion model from intervention-paired data with observed intervention targets, so that interventions correspond to well-defined operations in the exogenous manifold. We further provide an identifiability analysis that clarifies what aspects of this intrinsic representation are recoverable.

7 Conclusion

We proposed *CMT-Diff* to establish identifiable causal generation within diffusion models. By coupling a CNF with the PF-ODE, we constructed an exogenous manifold where interventions are realized as *CMT*. This mechanism is defined as vector translations along a learned causal basis. Our geometric formulation disentangles generative factors, enabling counterfactual control without compromising generative quality. Currently, the framework relies on intervention-paired data for manifold alignment. Future work will extend this approach to unsupervised settings, investigating whether intrinsic topological properties suffice to recover the causal basis.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean Government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); No. RS-2024-00457882, National AI Research Lab Project), and by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. RS-2025-16071925).

References

[Ahuja *et al.*, 2022] Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.

[Brehmer *et al.*, 2022] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[Burgess *et al.*, 2018] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[Chao *et al.*, 2024] Patrick Chao, Patrick Blöbaum, Sapan Kirit Patel, and Shiva Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *Transactions on Machine Learning Research*, 2024.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[Dominguez-Olmedo *et al.*, 2023] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, Georgios Arvanitidis, and Bernhard Schölkopf. On data manifolds entailed by structural causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8188–8201. PMLR, 23–29 Jul 2023.

[Eastwood and Williams, 2018] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

[Fatemi *et al.*, 2025] Pouria Fatemi, Ehsan Sharifian, and Mohammad Hossein Yassaee. A new approach to backtracking counterfactual explanations: A unified causal framework for efficient model interpretability. In *Forty-second International Conference on Machine Learning*, 2025.

[Grathwohl *et al.*, 2019] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[Hahm *et al.*, 2024] Jaehoon Hahm, Junho Lee, Sunghyun Kim, and Joonseok Lee. Isometric representation learning for disentangled latent space of diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17224–17245. PMLR, 21–27 Jul 2024.

[Hao *et al.*, 2024] Guang-Yuan Hao, Jiji Zhang, Biwei Huang, Hao Wang, and Kun Zhang. Natural counterfactuals with necessary backtracking. *Advances in Neural Information Processing Systems*, 37:14962–14995, 2024.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[Hinton *et al.*, 1984] Geoffrey E Hinton, James L McClelland, and David E Rumelhart. Distributed representations. Technical report, Carnegie-Mellon University, 1984.

[Huang *et al.*, 2024] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[Khemakhem *et al.*, 2020] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[Komanduri *et al.*, 2022] Aneesh Komanduri, Yongkai Wu, Wen Huang, Feng Chen, and Xintao Wu. Scm-vae: Learning identifiable causal representations via structural knowledge. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1014–1023, 2022.

[Komanduri *et al.*, 2023] Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. *arXiv preprint arXiv:2306.01213*, 2023.

[Komanduri *et al.*, 2024] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *Proceedings of the 27th European Conference on Artificial Intelligence*, 2024.

[Kwon *et al.*, 2023] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

[Lee, 2012] J. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2012.

[Lippe *et al.*, 2022] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[Lippe *et al.*, 2023] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

[Locatello *et al.*, 2020] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR, 2020.

[Park *et al.*, 2023] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[Peters *et al.*, 2017] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[Preechakul *et al.*, 2022] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.

[Sanchez and Tsaftaris, 2022] Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. In *Conference on Causal Learning and Reasoning*, pages 647–668. PMLR, 2022.

[Shuai *et al.*, 2024] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Von Kügelgen *et al.*, 2023] Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pages 177–196. PMLR, 2023.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Yang *et al.*, 2021] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9593–9602, June 2021.

[Yang *et al.*, 2023] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2023] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

[Zhou *et al.*, 2024] Zeyu Zhou, Ruqi Bai, Sean Kulinski, Murat Kocaoglu, and David I. Inouye. Towards characterizing domain counterfactuals for invertible latent causal models. In *The Twelfth International Conference on Learning Representations*, 2024.