Beyond the Black Box: Identifiable Interpretation and Control in Generative Models via Causal Minimality

Lingjing Kong*[1], Shaoan Xie*[1,2], Guangyi Chen[1,2], Yuewen Sun[1,2], Xiangchen Song[1], and Kun Zhang[1,2]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
*Equal contribution.

Abstract. Deep generative models, while revolutionizing image generation, largely operate as opaque "black boxes", hindering human understanding, control, and alignment. While methods like sparse autoencoders (SAEs) show remarkable empirical success, they often lack theoretical guarantees, risking subjective insights. Our primary objective is to establish a principled foundation for interpretable generative models. We demonstrate that the principle of causal minimality – favoring the simplest causal explanation – can endow the latent representations of generative models with clear causal interpretation and robust, component-wise identifiable control. We introduce a novel theoretical framework for hierarchical selection models, where higher-level concepts emerge from the constrained composition of lower-level variables, better capturing the complex dependencies in data generation. Under theoretically derived minimality conditions that manifest as sparsity constraints, we show that learned representations can be equivalent to the true latent variables of the data-generating process. Empirically, applying these constraints to text-to-image diffusion models allows us to extract their innate hierarchical concept graphs, offering fresh insights into their internal knowledge organization. Furthermore, these causally grounded concepts serve as levers for fine-grained model steering, paving the way for transparent, reliable systems.

1 Introduction

Deep generative models have transformed visual generation, with text-to-image diffusion models [11, 23, 52, 61, 68, 69] offering a prominent example. However, their escalating complexity and scale frequently cast them as opaque "black boxes" [53, 67]. This opacity presents a formidable barrier to genuine human understanding, severely curtails our ability to exert precise control over their behavior [19, 29, 65, 78], and complicates the crucial alignment with human values and intentions.

Although recent empirical tools, such as sparse autoencoders (SAEs) for diffusion models [10, 25, 34, 35, 73], offer avenues for probing these models, a

fundamental gap persists. Without rigorous theoretical underpinnings, interpretations derived from these methods risk being subjective or susceptible to human biases, rendering them potentially untrustworthy for risk-sensitive applications [31,50,63]. In this work, we directly tackle this critical challenge, seeking to establish a principled foundation for interpretable and controllable generative models.

Our investigation centers on two questions: Under what *theoretical conditions* can we reliably identify meaningful, interpretable latent concepts within the intricate architectures of modern generative models? And, crucially, what *actionable, theoretically-grounded insights* can empower us to advance both the interpretability and the controllability of these powerful systems?

Towards these goals, we identify the *causal minimality* [22,57,71] principle as the formal underpinning that connects widespread practices, such as enforcing sparsity, to the recovery of meaningful, interpretable concepts. This principle, advocating for the simplest causal model consistent with observations, allows for the identification of latent hierarchical concept structures. In our context, minimality translates to sparsity in the visual concept graphs of text-to-image (T2I) diffusion models [58, 59, 61]. Our findings indicate that imposing sparsity constraints on internal representations is instrumental for identifying intrinsic visual concepts. A cornerstone of our contribution is establishing the first identifiability results for *selection*-based [4, 6, 9, 14, 21, 72, 84, 88] hierarchical models. In such models, higher-level variables emerge as effects of compositions of lower-level variables, where higher-level variables control and select the configuration of lower-level ones. This fundamentally diverges from traditional hierarchical causal models [8, 56, 85], in which causal influence typically propagates from higher to lower levels. The selection model structure is particularly adept at capturing the intricate conditional dependencies among low-level features for forming coherent high-level concepts – it explains how specific arrangements of wheels, doors, and a roof constitute a recognizable "car", rather than a disjointed collection of parts. Traditional hierarchical models often neglect such intra-level dependencies by assuming no within-layer causal edges, as explicitly modeling them would yield overly dense graphs. The selection mechanism, in contrast, offers a simpler approach to this essential coordination. Its adherence to the minimality principle strongly favors it as a more accurate representation of the true model.

Despite the appeal, their identifiability has been underexplored. Prior research has largely centered on traditional hierarchical structures. Moreover, their techniques often rest on simplifying assumptions (e.g., linearity [1, 12, 24, 79] or achieve only subspace-level identifiability [37]). Such methods are generally inapplicable to the hierarchical selection models. Our framework establishes *component-wise* identifiability for continuous hierarchical selection models. Specifically, we demonstrate that under well-defined visual minimality conditions (Condition 1-iv), the learned representations are equivalent to the true latent variables of the underlying hierarchical process. This disentanglement of individual, atomic concepts is what affords significantly more nuanced interpretability and precise control in the resulting generative models.

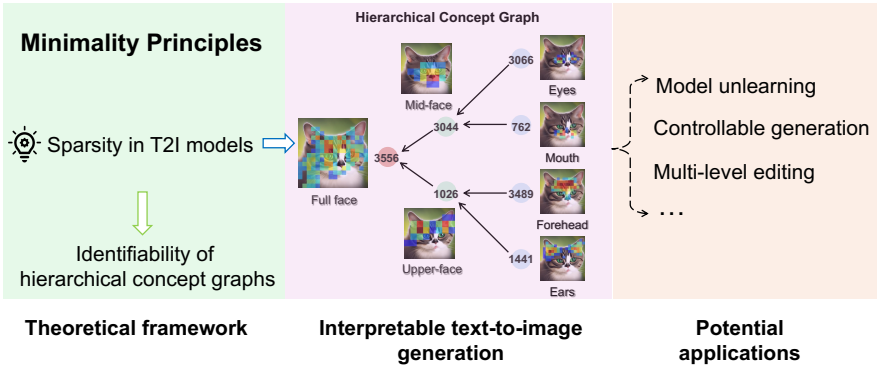

Fig. 1: Our causal minimality principle links sparsity in text-to-image models to identifiable hierarchical concept graphs, enabling interpretable generation and downstream control.

By applying the derived sparsity constraints to state-of-the-art text-to-image diffusion models, we successfully extract their innate hierarchical concept graphs (Figure 1). This not only illuminates their internal knowledge organization but also shows that causally-grounded concepts serve as highly effective levers for model steering. Our experiments illustrate key implications of our theorems and show how a principled, causal understanding can guide the application of established interpretation techniques.

2 Related Work

Hierarchical models. Complex real-world data distributions frequently exhibit inherent hierarchical structures among their underlying latent variables, a characteristic that has motivated extensive research. Initial explorations primarily focus on continuous latent variables with linear interactions [1, 12, 24, 79]. Other lines of work have centered on discrete latent variables; however, these approaches are often constrained in their applicability to continuous data modalities like images [8, 18, 36, 55, 85]. Furthermore, prevalent latent tree models, which connect variables via a single undirected path [8, 55, 85], risk oversimplifying the multifaceted relationships present in complex systems. Kong et al. [37] tackle nonlinear, continuous latent hierarchical models, but their framework, operating under rather opaque functional conditions, falls short of component-wise identifiability, thereby leaving room for concept entanglement. Our work distinctively investigates *selection* hierarchical models, contending that their structural properties yield a more faithful representation of latent concepts in natural data distributions. In these models, latent variables function as colliders, a significant departure from their role as confounders in the aforementioned prior art. This critical distinction renders existing identification techniques largely inapplicable. We provide *component-wise* identifiability for continuous hierarchical selection models.

Interpretability for generative models. Despite the remarkable advancements of generative models, their internal mechanisms often remain opaque. This presents a significant challenge to understanding and control. Considerable research has focused on obtaining interpretable features to enable more controllable generation. Early efforts center on analyzing the latent space of generative adversarial networks, e.g., [19, 66, 77]. Recently, sparse autoencoders (SAEs) have gained prominence for interpreting hidden representations in diffusion models. Surkov et al. [73] reveal interpretable features and specialization across diffusion model blocks. Other work trains SAEs with lightweight classifiers on diffusion model features [35] or steers generation away from undesirable visual attributes [25]. Our hierarchical approach is related to recent findings on the evolution of semantics during the diffusion process. It has been observed that high-level concepts, such as object shape and structure, tend to emerge in earlier, high-noise timesteps, while fine-grained, low-level details are synthesized in later, low-noise stages [49, 54, 74]. While these works provide valuable empirical validation of this phenomenon, our work offers a new perspective by framing these observations within a formal hierarchical, causal structure. We provide a theoretical foundation, rooted in causal minimality and selection models, to explain *how* these concepts compose and, crucially, *under what conditions* they can be provably identified. Our approach also relates to generative concept bottleneck models, which achieve interpretability by forcing predictions through a bottleneck layer of concepts [28, 40]. While these methods provide powerful intervention capabilities by design, our work differs by focusing on the discovery of the innate hierarchical and causal concept structure in the data. We provide the theoretical conditions for identifying these concepts component-wise, allowing us to then use this discovered graph for fine-grained multi-level interventions.

Decomposition-based interpretability. Our work is fundamentally distinct from post-hoc, decomposition-based interpretability methods, such as the prototype-matching approach [5]. This line of research, while pioneering, has known limitations (often stemming from its prototype-based implementation): its reliance on class-label supervision can lead to non-compositional, class-locked concepts [62], and its use of rigid patch-matching struggles with context and deformation [13, 81]. In contrast, our approach is *class/object agnostic* (similar to SAEs) and *context-sensitive*, learning from the raw generative data without class labels. Our approach learns compositional, shared concepts (e.g., a single "furry texture" from "cats" and "pandas") rather than rigid, class-specific prototypes. This enables the causal, interventional control (e.g., downstream tasks in Section 5.2) that prototype-matching cannot guarantee. Please find additional related work in Appendix A.

3 Deep Generative Models as Hierarchical Concept Models

Notations. We denote random variables with upper-case characters (e.g., X) and values with lower-case characters (e.g., x). We distinguish multidimensional

objects with bold fonts (e.g., $\mathbf{X}$) and refer to their dimensionality as $n(\cdot)$. We view multidimensional variables as *sets* when appropriate (e.g., $\mathbf{X}$ as $\{X_i\}_{i\in[n(\mathbf{X})]}$). Parents $\mathrm{Pa}(\cdot)$ and children $\mathrm{Ch}(\cdot)$ relations are defined based on the selection graph (Figure 2). If X has only one child Y, we refer to X as a pure parent of Y, i.e., $X \in \mathrm{PPa}(Y)$; if X has other children than Y, we refer to X as a hybrid parent of Y, i.e., $X \in \mathrm{HPa}(Y)$. We denote the set of natural numbers $\{1, \ldots, M\}$ as $[M]$. More background information is in Appendix C.

We denote the image as the continuous variable $\mathbf{X} \in \mathbb{R}^{n(\mathbf{X})}$ and text as the discrete variable $\mathbf{D} \in \mathbb{N}^{n(\mathbf{D})}$. Visual concepts are $\mathbf{Z} := [\mathbf{Z}_1, \cdots, \mathbf{Z}_{L_\mathrm{V}}]$, where L_V is the number of visual hierarchical levels and $\mathbf{Z}_l \in \mathbb{R}^{n(\mathbf{Z}_l)}$ are concepts at level l (Figure 2). The variables $\mathbf{D}$ encode the text prompt or other observed conditioning information. In contrast, the visual concepts ($\mathbf{Z}$) are continuous to represent rich visual details. $\mathbf{D}$ acts as a selection variable that governs the joint configuration of the continuous $\mathbf{Z}$ variables. For instance, the prompt concept "bicycle" selects for a coherent arrangement of continuous visual features representing wheels, a frame, and handlebars, rather than a random collection of those continuous parts.

Hierarchical processes and selection mechanisms. Our framework conceptualizes high-level concepts as emerging from or being effects of lower-level concepts. This is captured by a *selection mechanism* [4, 6, 9, 14, 21, 72, 84, 88], where variables $\mathbf{V}_l$ at a higher level of abstraction (smaller l) is determined by its constituent, more detailed components $\mathbf{V}_{l+1}$ (i.e., its "parents"). The selection function $g_{\mathbf{V}_l}$ maps these lower-level constituents to the higher-level concept:

$$\mathbf{V}_l := g_{\mathbf{V}_l}(\mathbf{V}_{l+1}). \tag{1}$$

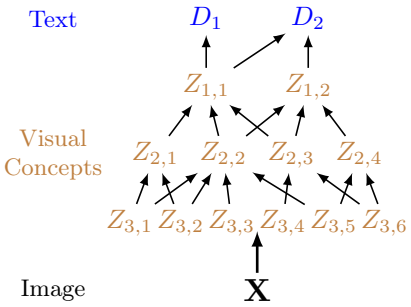

Fig. 2: A visual concept graph. We denote text conditioning as $\mathbf{D}$, visual concepts as $\mathbf{Z}$, and the image as $\mathbf{X}$. High-level concepts function as selection variables for low-level variables.

In other words, $\mathbf{V}_l$ is a selection variable over $\mathbf{V}_{l+1}$. In many natural data distributions of interest, we can only observe the data points for which the selection criterion is met, i.e., $\mathbf{V}_l$ only takes on a strict subset of its range Ω. Therefore, the distribution of $\mathbf{V}_{l+1}$ is always the conditional distribution $\mathbb{P}(\mathbf{V}_{l+1}|\mathbf{V}_l)$. This conditioning on $\mathbf{V}_l$ can induce dependencies among components in $\mathbf{V}_{l+1}$. For instance, if $V_{l+1,i} \to V_l \leftarrow V_{l+1,j}$, conditioning on V_l makes $V_{l+1,i}$ and $V_{l+1,j}$ dependent.

Under this formulation, one can leverage the inverse process of (1) to sample observable data (images, text), proceeding from higher-level abstract concepts to lower-level concrete details:

$$\mathbf{Z}_0 \sim \mathbb{P}(\mathbf{Z}_0), \quad \mathbf{Z}_l \sim \mathbb{P}(\mathbf{Z}_l|\mathbf{Z}_{l-1}), \quad l \in \{1, \ldots, L_\mathrm{V}+1\}, \tag{2}$$

where we denote $\mathbf{Z}_0 := \mathbf{D}$ and $\mathbf{Z}_{L_\mathrm{V}+1} := \mathbf{X}$. While (2) defines the generative pathway, the underlying structure is shaped by the selection principle of (1): the conditional distributions in (2) are implicitly learned if one has learned selection mechanisms in (1) and vice versa.

Why is this "selection" formulation? The "selection" perspective is critical for modeling how abstract concepts enforce coherence among their more concrete constituents. Consider generating an image of a "bicycle" (a high-level concept $\mathbf{Z}_l$). Its components – wheels, frame, handlebars (lower-level concepts $\mathbf{Z}_{l+1}$) – must not only be present but also be arranged in a specific, structurally sound configuration. Traditional hierarchical models [8,55,85] assume independent low-level concepts $Z_{l+1,i}$ given high-level concepts $\mathbf{Z}_l$ and stochastically sample these components, which could lead to unrealistic arrangements (e.g., wheels detached from the frame if the learned conditional is not perfect). Therefore, these models must additionally incorporate causal edges within each hierarchical level to capture this conditional dependency, resulting in highly dense causal graphs. In contrast, the selection model, by positing that $\mathbf{Z}_l$ is an effect of a specific configuration of $\mathbf{Z}_{l+1}$, emphasizes that the "bicycle" concept arises from a coherent selection and composition of its parts. This structured dependency, induced by the selection mechanism, yields a much simpler graphical model to describe the natural data distribution, thus preferred by the *minimality* principle.

Connections to text-to-image diffusion models. The iterative denoising process in diffusion aligns with our hierarchical data construction. These models involve a sequence of transformations $\{f_t\}_{t=1}^T$, parameterized by timestep t, that progressively restore a less noisy image $\mathbf{X}_t$ from a more corrupted version $\mathbf{X}_{t+1}$. As interpreted by Kong et al. [36], each f_{t+1} can be viewed as an autoencoder: it extracts a representation $\mathbf{Z}_{\mathcal{S}(t+1)}$ ($\mathcal{S}(t+1)$ indexes U-Net features associated with timestep $t+1$) from the noisy input $\mathbf{X}_{t+1}$, and uses this representation to produce the less noisy $\mathbf{X}_t$. In this view, representations $\mathbf{Z}_{\mathcal{S}(t+1)}$ from higher noise levels (larger t, where $\mathbf{X}_{t+1}$ is closer to pure noise) correspond to higher-level, more abstract concepts in our hierarchy (e.g., $\mathbf{Z}_l$ with smaller l), as fine-grained details are obscured by noise. Conversely, representations from lower noise levels (smaller t) capture more concrete details (e.g., $\mathbf{Z}_l$ with larger l). The diffusion model's step-wise refinement thus mirrors our hierarchical generation $\mathbb{P}(\mathbf{Z}_{l+1}|\mathbf{Z}_l)$, with the initial text prompt $\mathbf{D}$ typically guiding the most abstract visual concepts (e.g., $\mathbf{Z}_1 \sim \mathbb{P}(\mathbf{Z}_1|\mathbf{D})$, Figure 1). In our empirical analysis (Section 5), we explicitly map distinct diffusion timesteps to these hierarchical levels: high noise levels (e.g., $t = 899$) correspond to abstract concepts, while low noise levels (e.g., $t = 100$) correspond to fine-grained details.

Identifiability and interpretability. In light of the connection, a crucial question remains: are the internal representations learned by these models (e.g., U-Net or DiT features) truly reflective of the ground-truth concepts of the data, or are they merely effective for the generation task without being inherently interpretable and controllable? This motivates the need for *identifiability* guarantees that affirm the equivalence between the two worlds, which we present in Section 4.

4 Identifiable Representations under Causal Minimality

We first formally define our core theoretical principle, *causal minimality* [22, 57, 71]: Among all causal models that can explain the observed data, the true model is the simplest one. This principle is the key to our goal of identifiability (Definition 1). For visual concepts, causal minimality manifests as sparse connectivity in the causal graph (our minimality condition, 1-iv). Enforcing this sparsity is thus the practical mechanism that provides theoretical guarantees for identifiability.

For visual concepts, minimality manifests as a preference for *sparse graphical dependencies* within the latent hierarchy. This implies that concepts are formed through a limited set of direct causal influences, making the underlying structure easier to discern.

A key challenge we address is the identifiability of *hierarchical selection models*. In these models, higher-level concepts are effects of lower-level concepts. This contrasts with traditional hierarchical models where causality often flows from abstract to concrete, and where latent variables typically act as confounders [1, 8, 12, 18, 24, 36, 37, 55, 79, 85]. In our selection framework, latent variables act as colliders, rendering many existing identifiability results inapplicable. This distinction necessitates the novel theoretical development presented herein. Our goal is to achieve *component-wise identifiability*:

Definition 1 (Component-wise Identifiability). *Let* $\mathbf{Z}$ *and* $\hat{\mathbf{Z}}$ *be variables under two model specifications. We say that* $\mathbf{Z}$ *and* $\hat{\mathbf{Z}}$ *are* identified component-wise *if there exists a permutation* π *such that for each* $i \in [n(\mathbf{Z})]$, $\hat{Z}_i = h_i(Z_{\pi(i)})$ *where* h_i *is an invertible function.*

This strong form of identifiability ensures that each learned latent component $\hat{Z}_i$ corresponds to a single true latent component $Z_{\pi(i)}$. This is vital for unambiguous interpretation and targeted control. We assume the standard faithfulness condition [70], meaning the graphical model accurately reflects all conditional independence relations in the data.

In the following, we consider the identification of continuous latent visual concepts $\mathbf{Z}$.

Condition 1 (Visual Concept Identification Conditions)**.**

i ***Informativeness****: There exists a diffeomorphism* $g_l : (\mathbf{Z}_l, \boldsymbol{\epsilon}_l) \mapsto \mathbf{X}$ *for* $l \in [0, L]$*, where* $\boldsymbol{\epsilon}_l$ *denotes independent exogenous variables.*

ii ***Smooth Density****: The probability density function* $p(\mathbf{z}_{l+1}|\mathbf{z}_l)$ *is smooth for any* $l \in [L_{\mathrm{V}}]$.

iii ***Sufficient Variability****: For each* Z *and its parents* $\tilde{\mathbf{Z}} := \mathrm{Pa}(Z)$*, at any value* $\tilde{\mathbf{z}}$ *of* $\tilde{\mathbf{Z}}$*, there exist* $n(\tilde{\mathbf{Z}})+1$ *distinct values of* Z*, denoted as* $\{z^{(n)}\}_{n=0}^{n(\tilde{Z})}$*, such that the vectors* $\mathbf{w}(\tilde{\mathbf{z}}, z^{(n)}) - \mathbf{w}(\tilde{\mathbf{z}}, z^{(0)})$ *are linearly independent where*

$$\mathbf{w}(\tilde{\mathbf{z}}, z) = \Big(\frac{\partial \log p\left(\tilde{\mathbf{z}}|z\right)}{\partial \tilde{z}_1}, \ldots, \frac{\partial \log p\left(\tilde{\mathbf{z}}|z\right)}{\partial \tilde{z}_{n(\tilde{\mathbf{z}})}}\Big).$$

iv ***Sparse Connectivity (Minimality)****: For each parent concept $\tilde{Z}$, there exists a subset of its children $\mathbf{Z} \subseteq \mathrm{Ch}(\tilde{Z})$ such that their* only *common parent is $\tilde{Z}$, i.e.,*

$$\bigcap_{Z \in \mathbf{Z}} \mathrm{Pa}(Z) = \{\tilde{Z}\}.$$

Interpreting Condition 1. Condition 1-i ensures that the observed data $\mathbf{X}$ (e.g., an image) fully captures the information about the latent concepts $\mathbf{Z}_l$. This is a natural assumption as high-dimensional observations contain rich information. Condition 1-ii is a standard regularity assumption for analysis. Both are common in nonlinear ICA literature [26,27,32,33,37,76]. Condition 1-iii formalizes the idea that distinct lower-level concepts (e.g., "wheel," "door") respond in sufficiently distinct ways to changes in a shared higher-level concept (e.g., "car"), thus facilitating the identification of these lower-level concepts. Condition 1-iv is an instantiation of causal minimality for visual concepts. It posits that the causal graph of concepts is sparse: each concept has a unique "fingerprint" in terms of its connectivities. This sparsity is crucial for disentanglement [43–45,80,87] and is a less restrictive assumption than, for example, pure observed children for each latent variable [2,3,51]. This condition formalizes a core principle: concepts are learned through *comparison*. A concept is identifiable only if the data is rich enough to distinguish it from alternatives. For instance, if "Knight" and "Horse" always co-occur, they are learned as a fused concept; learning them separately requires data that breaks this correlation.

Theorem 2 (Visual Concept Identification). *Assume the process for visual concepts in* (2)*. Condition 1 describes the natural data distribution. For estimation, suppose a model specification $\boldsymbol{\theta}_{\mathrm{V}}$ satisfies Condition 1, and an alternative specification $\hat{\boldsymbol{\theta}}_{\mathrm{V}}$ satisfies Conditions 1-i and 1-ii, along with a sparsity constraint such that for corresponding $\hat{Z}$ and Z:*

$$n(\mathrm{Pa}(\hat{Z})) \leq n(\mathrm{Pa}(Z)), \tag{3}$$

then, if both models $\boldsymbol{\theta}_{\mathrm{V}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{V}}$ generate the same observed data distribution $\mathbb{P}(\mathbf{X})$, the latent visual concepts $\mathbf{Z}_l$ are component-wise identifiable for every level $l \in [L_{\mathrm{V}}]$.

Proof sketch for Theorem 2. The proof proceeds by identifying the hierarchical model level by level, from the top (most abstract concepts) $\mathbf{Z}_1$ downwards to $\mathbf{Z}_{L_{\mathrm{V}}}$. 1) The paired text data $\mathbf{D}$ acts as an auxiliary variable, providing diverse "influences" on the top-level $\mathbf{Z}_1$. Condition 1-iii ensures these interventions have distinguishable effects. Analogous to techniques in nonlinear ICA [26, 27, 39], each component D allows the identification of the subspace of $\mathbf{Z}_1$ variables it influences. 2) With these subspaces identified, one can identify the intersection of these subspaces [38,76,82]. Therefore, if the graphical structure is sufficiently sparse, as specified in Condition 1-iv, one can identify the top-level latent variable $\mathbf{Z}_1$ component-wise. 3) Once $\mathbf{Z}_1$ is identified, its components can serve as the auxiliary variables to identify the next level, $\mathbf{Z}_2$. This process is repeated iteratively down the hierarchy, identifying $\mathbf{Z}_l$ using the already identified $\mathbf{Z}_{l-1}$.

Implications for text-to-image diffusion models. Theorem 2 underscores that the *sparsity constraint* (3) *is pivotal for identifying true visual concepts*. In practice, this constraint is instantiated through a two-step process: 1) Level-specific concept learning: We train K-sparse SAEs on features at the specific timesteps defined in Section 3. This approximates the sparsity condition required by Theorem 2. 2) Cross-level causal discovery: We then apply causal discovery algorithms (e.g., PC [70]) across these sparse features to construct the hierarchical graph, validating that the learned representations align with the theoretical identification guarantees.

5 Experiments

Evaluation design and objectives. We design our experiments to validate our theoretical framework in two ways. In Section 5.1, we provide a direct empirical test of our theory: we apply the sparsity constraints derived from causal minimality (Condition 1) and show that we can, as predicted, extract a meaningful and interpretable hierarchical concept graph. In Section 5.2, we demonstrate the utility of these identified concepts. If our concepts are truly component-wise identifiable (Definition 1), they should be individually controllable. We test this via a suite of challenging downstream tasks—including model unlearning, controllable image generation, and multi-level editing. For example, we compare against state-of-the-art unlearning methods to rigorously benchmark our concept removal capabilities. Our objective is to show that our theory not only finds interpretable concepts but also provides a practical mechanism for fine-grained, reliable model control. More detailed settings for each experiment are provided in their respective subsections.

Hierarchical causal analysis. Our theoretical framework motivates an empirical analysis that differs from standard interpretability approaches. Following the framework established in Sections 3 and 4, we apply our two-step identification process to Stable Diffusion (SD) 1.4 [60] and Flux.1-Schnell [42] (Appendix E). We analyze feature representations at the previously defined timesteps (899, 500, and 100) to extract and verify the hierarchical concept graph.

Benefits. This hierarchical perspective provides two main benefits. First, it enables compositional editing. For a complex object like “a textured tree stump”, our analysis can distinguish the ”stump” (a mid-level concept) from its “texture” (a low-level one), allowing for independent steering. This is a fine-grained control challenging for non-hierarchical methods that tend to learn entangled features. Second, it allows for targeted intervention. By identifying a concept’s level, we can inject a steered feature back into the diffusion process only at its corresponding timestep, which helps in reducing the unwanted artifacts that can arise from applying steering globally across all timesteps (see Figure 5). More details in Appendix D.

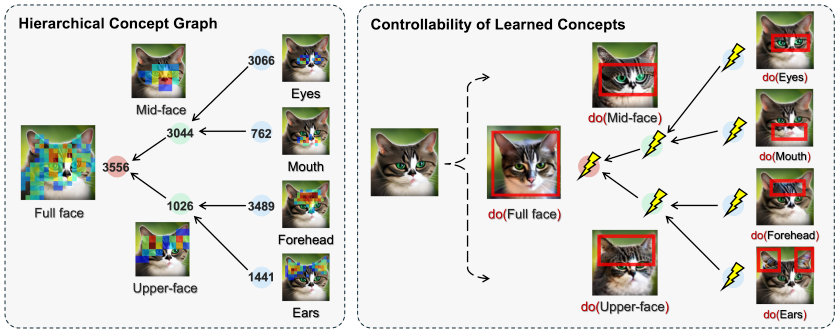

Fig. 3: Examples of hierarchical concept graphs for text-to-image models. Our method successfully recovers meaningful hierarchical structures, where each node encodes distinct semantic concepts. On the right, we demonstrate feature steering, where manipulating individual nodes leads to changes in the output that align with their position in the hierarchy. Intervening on a high-level concept in the learned graph ("Full face") alters the cat's entire facial structure and fur pattern. In contrast, intervening on a learned lower-level concept (e.g., "Eye") produces a much more localized edit, changing only the shape and color of the eyes while leaving the rest of the face intact. More examples in Appendix E.

Table 1: Model unlearning comparisons. Our method delivers competitive results on unlearning tasks without compromising standard text-to-image generation. See Appendix C for details.

Method	**I2P** ↓	**RING-A-BELL** ↓				**P4D** ↓	**UATK** ↓	**COCO**	
		K77	K38	K16	AVG			FID ↓	CLIP ↑
SD 1.4 [61]	17.8	85.26	87.37	93.68	88.10	98.70	69.70	16.71	**31.3**
ESD [15]	2.87	20.00	29.47	35.79	28.42	15.49	2.87	18.18	30.2
SA [20]	2.81	63.15	56.84	56.84	58.94	12.68	2.81	25.80	29.7
CA [41]	1.04	86.32	91.69	94.26	90.76	5.63	1.04	24.12	30.1
MACE [48]	1.51	2.10	**0.00**	**0.00**	**0.70**	2.82	1.51	**16.80**	28.7
UCE [16]	0.87	10.52	9.47	12.61	10.87	9.86	0.87	17.99	30.2
RECE [17]	0.72	5.26	4.21	5.26	4.91	5.63	0.72	17.74	30.2
SDID [46]	3.77	94.74	95.79	90.53	93.68	69.54	30.99	22.16	31.1
SLD-MAX [64]	1.74	23.16	32.63	42.11	32.63	9.14	2.44	28.75	28.4
SLD-STRONG [64]	2.28	56.84	64.21	61.05	60.70	33.10	3.10	24.40	29.1
SLD-MEDIUM [64]	3.95	92.63	88.42	91.05	90.70	24.00	1.98	21.17	29.8
SD1.4-NegPrompt [61]	0.74	17.89	40.42	34.74	31.68	10.00	1.46	18.33	30.1
SAFREE [83]	1.45	35.78	47.36	55.78	46.31	10.56	1.45	19.32	30.1
TRASCE [30]	0.45	1.05	2.10	2.10	1.75	3.97	**0.70**	17.41	29.9
ConceptSteer [34]	0.36	3.16	8.42	9.47	7.02	1.99	2.11	18.67	30.8
Ours	**0.26** (0.06)	**2.10** (0.86)	**0.00** (0.00)	**1.41** (0.99)	1.17	**0.66**	2.11	17.02	**31.3**

5.1 Interpretability Analysis

Hierarchical concept graph. Figure 3 illustrates a hierarchical graph learned through our approach (more in Appendix E). On the left, we display activation maps of different SAE features. Brown nodes (SAE nodes trained on timestep 899) capture high-level features, such as node 3556 representing an entire cat

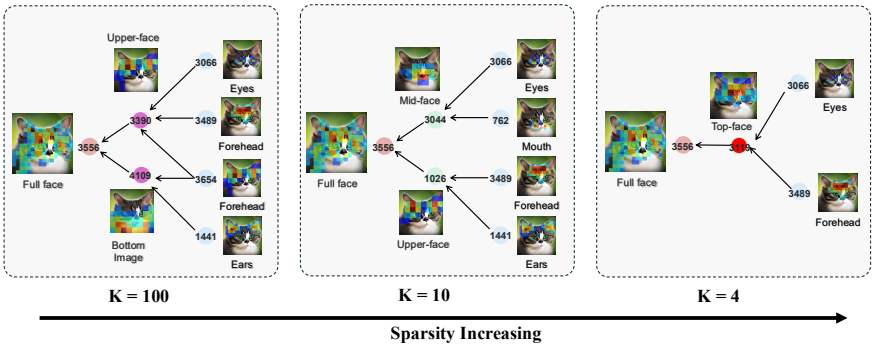

Fig. 4: Ablation studies on the sparsity constraint. We control feature sparsity at timestep 500 by varying the number of features (K) in SAE. Without enforcing sparsity, the resulting concepts tend to be dense, and the features are less interpretable. Conversely, higher sparsity leads to a more interpretable, sparser graph. However, when sparsity becomes too high, the resulting graph may become overly sparse and fail to adequately capture the generation of the cat face.

face. Green nodes (timestep 500) reflect mid-level features, like node 3044 capturing the central face. Blue nodes (timestep 100) capture fine details—node 3066 activates on the eyes and node 762 on the mouth. This demonstrates a clear progression from coarse to fine-grained concepts across timesteps. To thoroughly examine the existence of the hierarchical concept graph, we conduct two complementary experiments demonstrating that activations at higher timesteps capture more global semantics, while those at lower timesteps capture more localized details. First, we quantify the spatial spread of activations across timesteps. For each SAE, we compute attribution maps for its top feature indices. Given an SAE feature of shape $64 \times 64 \times 5120$, we compute a 64×64 attribution map. Applying a 0.1 threshold yields a binary attribution map, from which we measure the proportion of activated pixels. Across 1,000 samples, approximately 280, 630, 880, and 1,400 unique concepts are activated for $K = 1, 3, 5$, and 10, respectively. Activations at timestep 899 influence a larger spatial area, indicating that higher timesteps capture more global, distributed concepts. Second, we generate images from 10,000 COCO prompts and deactivate the top-1 SAE activation at each timestep. Comparing the modified generations with the originals shows that deactivations at noisier timesteps cause substantial, global changes, while those at less noisy timesteps produce localized effects. These results confirm that features at different noise levels encode distinct abstraction levels, supporting the hierarchical concept graph.

Concept steering in hierarchical graphs. We conduct concept steering using our discovered features, as shown on the right side of Figure 3 (more in Appendix E). Given a model intermediate feature x, the SAE encoder E and decoder D are trained to reconstruct x. To steer a specific concept, we obtain the latent representation $z = E(x)$, and extract the steering vector v corresponding

Table 2: Quantitative analyses across different noise levels. (a) Spatial activation spread: average proportion of pixels influenced by the top-k SAE activations. Higher timesteps affect a larger spatial area, indicating that SAEs at noisier steps capture more global, distributed concepts. (b) SAE-deactivated generation comparison: similarity metrics between original and SAE-deactivated images. Deactivation at higher timesteps produces greater perceptual and semantic changes, supporting the presence of a hierarchical organization of concepts across timesteps.

	(a) Spatial activation spread				**(b) SAE-deactivation comparison**			
Timestep	**Top1**	**Top3**	**Top5**	**Top10**	**L1**	**LPIPS**	**CLIP**	**DINO**
100	0.27	0.21	0.19	0.15	0.004	0.002	0.999	0.999
500	0.30	0.25	0.21	0.17	0.013	0.020	0.995	0.993
899	0.53	0.41	0.33	0.24	0.070	0.220	0.948	0.903

Table 3: Controllable image generation results. Our method achieves the best CLIP-I metric, demonstrating greater fidelity to the input images, while reliably executing the target edits.

Metric	**SD 1.4**	**SD1.4 (SAE w/o hier.)**	**SD1.4 (Ours)**
Add tabby pattern – CLIP-I ↓	0.91 ± 0.05	0.83 ± 0.07	**0.93 ± 0.04**
Add tabby pattern – CLIP-T ↑	0.27 ± 0.00	**0.28 ± 0.02**	**0.28 ± 0.01**
Add mountains – CLIP-I ↓	0.84 ± 0.06	0.83 ± 0.04	**0.91 ± 0.03**
Add mountains – CLIP-T ↑	**0.33 ± 0.01**	0.32 ± 0.01	**0.33 ± 0.01**
Replace rock w/ stump – CLIP-I ↓	0.93 ± 0.02	0.95 ± 0.02	**0.96 ± 0.02**
Replace rock w/ stump – CLIP-T ↑	**0.31 ± 0.01**	0.29 ± 0.01	**0.31 ± 0.01**

to the desired feature. We then modify the original feature to create a steered version $x' = x + \lambda D(v)$, where λ modulates the strength. By feeding the steered x' back into the diffusion process at the same timestep, we generate images that reflect the influence of the selected concept. For example, steering node 3556 – associated with the entire face of a cat – results in a significantly altered cat face. Steering the green node 1026 modifies only the upper part of the face, illustrating that it encodes localized information specific to that region.

Ablation. As established in the theoretical framework, sparsity is crucial for identifiability. To empirically validate this, we visualize the resulting causal graphs under varying levels of sparsity, as shown in Figure 4 (more in Appendix E). When sparsity is not enforced, the resulting graph becomes overly dense, making it difficult to interpret and diminishing its semantic clarity. Conversely, imposing excessive sparsity leads to an overly pruned graph that lacks sufficient structure to meaningfully explain the generation process, such as in the case of the cat image. These observations highlight the importance of balancing sparsity to preserve interpretability while maintaining explanatory power.

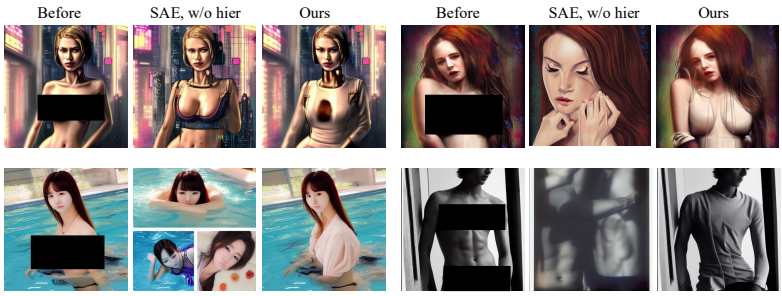

Fig. 5: Generated samples with P4D prompts [7]. The Stable Diffusion model is vulnerable to the prompts in the p4d dataset, producing unsafe images. When the hierarchical relationship across timesteps is not considered, negative steering with SAE results in drastic changes to the output. In contrast, our method learns to apply modifications to the nudity feature at a suitable timestep without additional distortions.

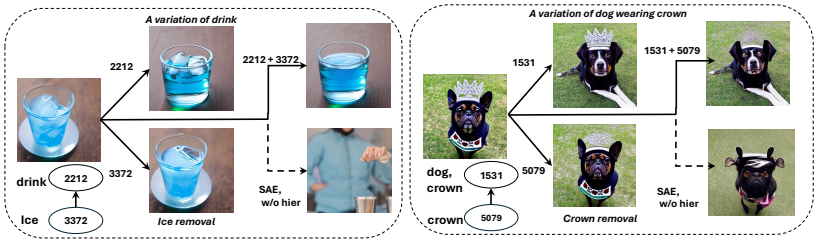

Fig. 6: Examples of multi-level editing (best viewed with zoom). High-level node 2212 contains all information about the cup, while mid-level node 3372 focuses primarily on the ice cubes. Similarly, high-level node 1531 encompasses all information about the dog (including the crown), and mid-level node 5079 is dedicated to the crown. By modeling hierarchical relationships, we can perform edits that are often difficult to achieve with a single-layer edit. For instance, if we want to generate a variation of the cup while removing the ice cubes, we can apply feature steering on high-level node 2212 to create a new version of the cup, and simultaneously apply negative feature steering on mid-level node 3372 to remove the ice cubes.

5.2 Downstream Tasks

Thanks to our theoretical framework, we can naturally perform a range of image generation and editing tasks, including model unlearning, controllable image generation, and multi-level editing.

Model unlearning. We provide quantitative results of model unlearning on four benchmark datasets: IP2P [64], three splits of RING-A-BELL [75], P4D [7], and UnlearnDiffATK [86]. These benchmarks focus on removing nudity-related concepts, and we report the accuracy of a pretrained nudity detector. Our method achieves competitive concept suppression while preserving standard text-to-image generation quality. In addition, to assess whether our method preserves general text-to-image capability, we apply feature steering on normal prompts from

MSCOCO [47]. The 10K results, reflected in low FID and high CLIP scores, demonstrate that our method successfully identifies and removes nudity concepts without affecting unrelated concepts. Appendix E also reports complementary VLM under-erasure and retention metrics and style-removal results.

Controllable image generation. We also evaluate controllable image generation on three editing tasks: adding tabby patterns to cat faces, adding mountains to landscape images, and replacing rocks with textured tree stumps. As shown in Table 3, our method achieves superior results compared to both the standard text-guided model and SAE without hierarchical modeling.

Multi-level image editing. A key advantage of the hierarchical concept graph is that it can combine nodes across different levels for fine-grained image editing. In Figure 6, to obtain a new drink without ice (while preserving the background), we can apply multi-level editing by steering features at both high-level node 2212 and mid-level node 3372 simultaneously. Without such hierarchical relationship modeling, conventional methods struggle to produce this combination, which can result in undesired changes such as the drink being replaced by a person or the dog's background.

6 Conclusion

In this work, we present a theoretical framework using causal minimality for identifying latent concepts in hierarchical selection models. We prove that generative model representations can map to true latent variables. Empirically, applying these constraints to text-to-image diffusion models enables extracting meaningful hierarchical concept graphs, enhancing interpretability and grounded control.

Limitations: Our identifiability results are derived under standard regularity/minimality assumptions. These conditions are consistent with the inductive biases we encourage and are supported by our empirical findings. Extending the guarantees to broader forms of misspecification remains an important direction for future work.

Acknowledgements

We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, MBZUAI-WIS Joint Program, and the Al Deira Causal Education project.

References

1. Anandkumar, A., Hsu, D., Javanmard, A., Kakade, S.: Learning linear bayesian networks with latent variables. In: International Conference on Machine Learning. pp. 249–257. PMLR (2013)

2. Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: International conference on machine learning. pp. 280–288. PMLR (2013)
3. Arora, S., Ge, R., Moitra, A.: Learning topic models–going beyond svd. In: 2012 IEEE 53rd annual symposium on foundations of computer science. pp. 1–10. IEEE (2012)
4. Bareinboim, E., Tian, J., Pearl, J.: Recovering from selection bias in causal and statistical inference. Probabilistic and causal inference: The works of Judea Pearl pp. 433–450 (2022)
5. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
6. Chen, L., Zoeter, O., Mooij, J.M.: Modeling latent selection with structural causal models. arXiv preprint arXiv:2401.06925 (2024)
7. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. arXiv preprint arXiv:2309.06135 (2023)
8. Choi, M.J., Tan, V.Y., Anandkumar, A., Willsky, A.S.: Learning latent tree graphical models. Journal of Machine Learning Research **12**, 1771–1812 (2011)
9. Correa, J.D., Tian, J., Bareinboim, E.: Identification of causal effects in the presence of selection bias. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 2744–2751 (2019)
10. Cywiński, B., Deja, K.: Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. arXiv preprint arXiv:2501.18052 (2025)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021) (2021)
12. Dong, X., Huang, B., Ng, I., Song, X., Zheng, Y., Jin, S., Legaspi, R., Spirtes, P., Zhang, K.: A versatile causal discovery framework to allow causally-related hidden variables. In: The Twelfth International Conference on Learning Representations (2023)
13. Donnelly, J., Barnett, A.J., Chen, C.: Deformable protopnet: An interpretable image classifier using deformable prototypes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10265–10275 (2022)
14. Forré, P., Mooij, J.M.: Causal calculus in the presence of cycles, latent confounders and selection bias. In: Uncertainty in Artificial Intelligence. pp. 71–80. PMLR (2020)
15. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2426–2436 (2023)
16. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
17. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models. In: European Conference on Computer Vision. pp. 73–88. Springer (2024)
18. Gu, Y., Dunson, D.B.: Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. In: Journal of the Royal Statistical Society Series B: Statistical Methodology (2023)
19. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. Advances in neural information processing systems **33**, 9841–9850 (2020)

20. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems **36**, 17170–17194 (2023)
21. Hernán, M.A., Hernández-Díaz, S., Robins, J.M.: A structural approach to selection bias. Epidemiology **15**(5), 615–625 (2004)
22. Hitchcock, C.: Probabilistic Causation. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2021 edn. (2021)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (2020)
24. Huang, B., Low, C.J.H., Xie, F., Glymour, C., Zhang, K.: Latent hierarchical causal structure discovery with rank constraints. Advances in Neural Information Processing Systems **35**, 5549–5561 (2022)
25. Huang, V.S.J., Zhuo, L., Xin, Y., Wang, Z., Gao, P., Li, H.: Tide: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. arXiv preprint arXiv:2503.07050 (2025)
26. Hyvarinen, A., Morioka, H.: Unsupervised feature extraction by time-contrastive learning and nonlinear ica. Advances in neural information processing systems **29** (2016)
27. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ica using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 859–868. PMLR (2019)
28. Ismail, A.A., Adebayo, J., Bravo, H.C., Ra, S., Cho, K.: Concept bottleneck generative models. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=L9U5MJJleF`
29. Jahanian, A., Chai, L., Isola, P.: On the" steerability" of generative adversarial networks. In: International Conference on Learning Representations (2019)
30. Jain, A., Kobayashi, Y., Shibuya, T., Takida, Y., Memon, N., Togelius, J., Mitsufuji, Y.: Trasce: Trajectory steering for concept erasure. arXiv preprint arXiv:2412.07658 (2024)
31. Kaddour, J., Lynch, A., Liu, Q., Kusner, M.J., Silva, R.: Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475 (2022)
32. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics. pp. 2207–2217. PMLR (2020)
33. Khemakhem, I., Monti, R., Kingma, D., Hyvarinen, A.: Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. Advances in Neural Information Processing Systems **33**, 12768–12778 (2020)
34. Kim, D., Ghadiyaram, D.: Concept steerers: Leveraging k-sparse autoencoders for controllable generations. arXiv preprint arXiv:2501.19066 (2025)
35. Kim, D., Thomas, X., Ghadiyaram, D.: Revelio: Interpreting and leveraging semantic information in diffusion models. arXiv preprint arXiv:2411.16725 (2024)
36. Kong, L., Chen, G., Huang, B., Xing, E.P., Chi, Y., Zhang, K.: Learning discrete concepts in latent hierarchical models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), `https://openreview.net/forum?id=bO5bUxvH6m`
37. Kong, L., Huang, B., Xie, F., Xing, E., Chi, Y., Zhang, K.: Identification of nonlinear latent hierarchical models. Advances in Neural Information Processing Systems **36** (2023)

38. Kong, L., Ma, M.Q., Chen, G., Xing, E.P., Chi, Y., Morency, L.P., Zhang, K.: Understanding masked autoencoders via hierarchical latent variable models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7918–7928 (June 2023)
39. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., Zhang, K.: Partial disentanglement for domain adaptation. In: International Conference on Machine Learning. pp. 11455–11472. PMLR (2022)
40. Kulkarni, A., Yan, G., Sun, C.E., Oikarinen, T., Weng, T.W.: Interpretable generative models through post-hoc concept bottlenecks. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8162–8171 (2025)
41. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
42. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
43. Lachapelle, S., Deleu, T., Mahajan, D., Mitliagkas, I., Bengio, Y., Lacoste-Julien, S., Bertrand, Q.: Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. arXiv preprint arXiv:2211.14666 (2022)
44. Lachapelle, S., López, P.R., Sharma, Y., Everett, K., Priol, R.L., Lacoste, A., Lacoste-Julien, S.: Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. arXiv preprint arXiv:2401.04890 (2024)
45. Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K.E., Le Priol, R., Lacoste, A., Lacoste-Julien, S.: Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In: Conference on Causal Learning and Reasoning. pp. 428–484. PMLR (2022)
46. Li, H., Shen, C., Torr, P., Tresp, V., Gu, J.: Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12006–12016 (2024)
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
48. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
49. Mahajan, S., Rahman, T., Yi, K.M., Sigal, L.: Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6808–6817 (2024)
50. Moran, G.E., Aragam, B.: Towards interpretable deep generative models via causal representation learning (2025)
51. Moran, G.E., Sridhar, D., Wang, Y., Blei, D.M.: Identifiable variational autoencoders via sparse decoding. arXiv preprint arXiv:2110.10804 (2021)
52. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning (ICML 2021) (2021)
53. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S.: Zoom in: An introduction to circuits. Distill (2020). `https://doi.org/10.23915/distill.00024.001`, https://distill.pub/2020/circuits/zoom-in

54. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23051–23061 (2023)
55. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
56. Pearl, J.: Causality. Cambridge university press (2009)
57. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
58. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2022) (2022)
59. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
61. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022) (2022)
62. Rymarczyk, D., Struski, L., Tabor, J., Zieliński, B.: Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: KDD '21. p. 1420–1430. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447548.3467245, https://doi.org/10.1145/3447548.3467245
63. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
64. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
65. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020)
66. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence **44**(4), 2004–2018 (2020)
67. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810 (2017)
68. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning (ICML 2015) (2015)
69. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR 2022) (2022)
70. Spirtes, P., Glymour, C., Scheines, R.: Causation, prediction, and search. MIT press (2001)
71. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search. MIT press (2000)

72. Spirtes, P., Meek, C., Richardson, T.: Causal inference in the presence of latent variables and selection bias. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. pp. 499–506 (1995)
73. Surkov, V., Wendler, C., Terekhov, M., Deschenaux, J., West, R., Gulcehre, C.: Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders. arXiv preprint arXiv:2410.22366 (2024)
74. Tinaz, B., Fabian, Z., Soltanolkotabi, M.: Emergence and evolution of interpretable concepts in diffusion models. arXiv preprint arXiv:2504.15473 (2025)
75. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? arXiv preprint arXiv:2310.10012 (2023)
76. Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F.: Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems **34**, 16451–16467 (2021)
77. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: International conference on machine learning. pp. 9786–9796. PMLR (2020)
78. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12863–12872 (2021)
79. Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., Zhang, K.: Identification of linear non-gaussian latent hierarchical structure. In: International Conference on Machine Learning. pp. 24370–24387. PMLR (2022)
80. Xu, D., Yao, D., Lachapelle, S., Taslakian, P., Von Kügelgen, J., Locatello, F., Magliacane, S.: A sparsity principle for partially observable causal representation learning. arXiv preprint arXiv:2403.08335 (2024)
81. Xue, M., Huang, Q., Zhang, H., Hu, J., Song, J., Song, M., Jin, C.: Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. In: IJCAI (2024)
82. Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., Locatello, F.: Multi-view causal representation learning with partial observability. In: The Twelfth International Conference on Learning Representations (2023)
83. Yoon, J., Yu, S., Patil, V., Yao, H., Bansal, M.: Safree: Training-free and adaptive guard for safe text-to-image and video generation. arXiv preprint arXiv:2410.12761 (2024)
84. Zhang, J.: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence **172**(16-17), 1873–1896 (2008)
85. Zhang, N.L.: Hierarchical latent class models for cluster analysis. The Journal of Machine Learning Research **5**, 697–723 (2004)
86. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In: European Conference on Computer Vision. pp. 385–403. Springer (2024)
87. Zheng, Y., Ng, I., Zhang, K.: On the identifiability of nonlinear ica: Sparsity and beyond. arXiv preprint arXiv:2206.07751 (2022)
88. Zheng, Y., Tang, Z., Qiu, Y., Schölkopf, B., Zhang, K.: Detecting and identifying selection structure in sequential data. In: International Conference on Machine Learning. pp. 61498–61525. PMLR (2024)