Intervening to learn and compose disentangled representations

Alex Markham[1] Jeri A. Chang[2] Isaac Hirsch[2] Liam Solus[3] Bryon Aragam[2]

[1]Dept. of Mathematical Sciences, University of Copenhagen
[2]Booth School of Business, University of Chicago
[3]Department of Mathematics, KTH Royal Institute of Technology

Abstract

In causal representation learning, it is commonly believed that in order to learn useful latent structure, we face a fundamental tension between expressivity and structure. In this paper we challenge this view by proposing a new approach to training arbitrarily expressive generative models that also learn disentangled latent structure that enables multi-concept interventions and out-of-distribution (OOD) composition. This is accomplished by adding a simple decoder-only module to the head of an existing decoder block that can be arbitrarily complex. The module learns to process concept information by implicitly inverting linear representations from an encoder. Inspired by the notion of intervention in causal graphical models, our module selectively modifies its architecture during training, allowing it to learn a compact joint model over different contexts. We show how adding this module leads to disentangled representations that can be composed for OOD generation. To further validate our proposed approach, we show how our module approximates an identifiable concept model by establishing an identifiability result that extends existing work on identifying structured representations in nonlinear models.

1 INTRODUCTION

Generative models have transformed information processing and demonstrated remarkable capacities for creativity in a variety of tasks ranging from vision to language to audio. The success of these models has been largely driven by modular, differentiable architectures based on deep neural networks that learn useful representations for downstream tasks. Recent years have seen increased interest in understanding and exploring the representations produced by these models through evolving lines of work on structured representation learning, identifiability and interpretability, disentanglement, and causal generative models. This work is motivated in part by the desire to produce performant generative models that also capture meaningful, semantic latent spaces that enable out-of-distribution generation under perturbations.

A key conceptual driver of this line of work is the tradeoff between flexibility and structure, or expressivity and interpretability: It is widely believed that to learn structured, interpretable representations, model capacity must be constrained, sacrificing flexibility and expressivity. As described in Section 3, this intuition is supported by a growing body of theoretical work on nonlinear ICA, disentanglement, and causal representation learning, for example. On the practical side, methods that have been developed to learn structured latent spaces tend to be bespoke to specific data types and models, and typically impose significant limitations on the flexibility and expressivity of the underlying models. Moreover, the most successful methods for learning structured representations typically impose fixed, known structure *a priori*, as opposed to learning this structure from the data.

At the same time, there is a growing body of work that suggests generative models already learn surprisingly structured latent spaces (see Section 3 for more discussion). This in turn suggests that existing models are already “close” to capturing the desired structure, and perhaps only small modifications are needed. Since we already have performant models that achieve state-of-the-art results in generation and discrimination, do we need to re-invent the wheel to achieve these goals? Our hypothesis is that we should be able to leverage the expressiveness of these models to build new models that learn latent structure from scratch. We emphasize up front that our goal is not to explain or interpolate the latent space of a pre-trained model, but rather to leverage known architectures to train a *new* model end-to-end from

Full paper: https://arxiv.org/abs/2507.04754.

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

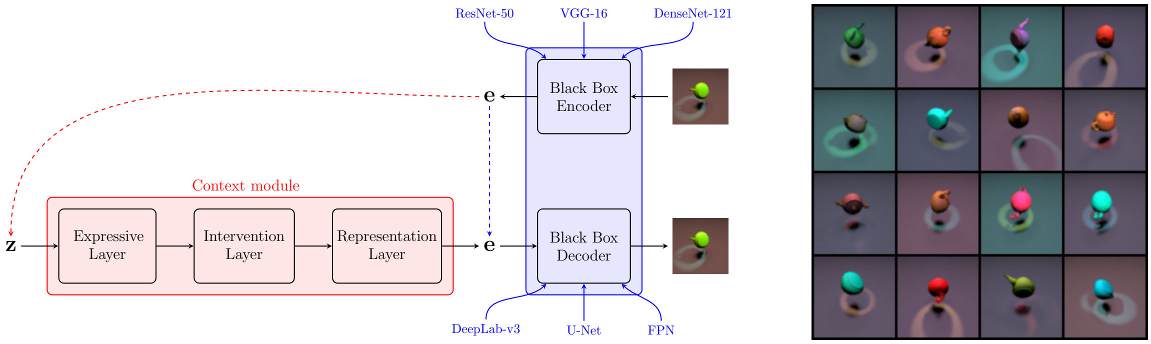


Figure 1: Overview of our approach. Given a black-box encoder-decoder architecture (blue), we propose to append a *context module* (red) to the head of the decoder. (left) Instead of passing the output of the encoder directly to the decoder (blue box + blue arrow), the embeddings $\mathbf{e}$ are passed through the context module, consisting of three distinct layers. The output of this module is then passed into the decoder. (right) The model learns to compose different concepts OOD, in this case object and background color, neither of which appear together in the training data.

scratch.

In this paper, we adopt this perspective: We start with a black-box encoder-decoder architecture, upon which we make no assumptions, and then augment this model to learn structured representations. The idea is that the black-box is an architecture that is already known to perform well on downstream tasks such as sampling and prediction, and so is flexible enough to capture complex patterns in data. We then introduce a modular, end-to-end differentiable architecture for learning structured representations, augmenting the original model with a simple decoder-only module that retains its existing capabilities while enabling concept identification and intervention as well as composing together multiple concept interventions for out-of-distribution (OOD) generation. The resulting latent structure is learned fully end-to-end, with no fixed structure imposed *a priori*. The motivation is to provide a framework for taking existing performant architectures and to train a new model that performs just as well, but with the added benefit of learning structured representations without imposing specific prior knowledge or structure *a priori*. These representations may be useful for interpretability, downstream tasks, or other purposes. Since the model, unlike previous approaches, imposes no latent structure a priori, we theoretically validate the approach with a novel identifiability result for structured representations recovered by the model.

To explore the implications of this modification, we run experiments including OOD generation. An attractive feature of our approach is that it makes running carefully controlled experiments and ablations particularly simple: We can carefully control hyperparameters across different models, ensuring that any differences in performance are attributable to specifically controlled architectural modifications. As a result, we contribute to a growing body of empirical work on the OOD capabilities of generative models by introducing a new set of ablations that can be used to guide the development of new methods.

Contributions Our main contributions can be summarized as follows:

1. We introduce a simple module that can be attached to the head of an existing decoder block to learn concept representations by splitting each concept into a tensor slice, where each slice represents an interventional context in a reduced form structural equation model (SEM). This module can be attached to the head of any decoder block and trained end-to-end.
2. We conduct experiments on disentanglement and OOD generation on several real datasets as well as controlled simulations. To this end, we introduce a simple simulated visual environment for testing disentanglement and OOD generation. This allows us to carefully test different ablations under controlled settings.
3. To illustrate the adaptivity of our module to different architectures, we include experiments with several different VAE architectures including NVAE [71].

As a matter of independent interest, we prove an identifiability result under concept interventions, and discuss how the proposed architecture can be interpreted as an approximation to this model. Full technical proofs are omitted and have been deferred to the full version of the paper.

2 PRELIMINARIES

Let $\mathbf{x}$ denote the observations, e.g., pixels in an image, and $\mathbf{z}$ denote hidden variables, e.g. latent variables that are to be inferred from the pixels. We are interested in training generative models with an encoder-decoder architecture, such as a variational autoencoder (VAE). A typical generative model consists of a decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ and an encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$.

After specifying a prior $p_\theta(\mathbf{z})$, this defines a likelihood by

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z}) \,\mathrm{d}\mathbf{z}. \tag{1}$$

Both the decoder and encoder are specified by deep neural networks that will be trained end-to-end via standard techniques [29, 58, 57]. These networks are often chosen to capture implicit assumptions about the structure of the latent space, such as hierarchical structure, graphical structure, equivariance, etc. In this work, we focus on VAEs since they are often use to represent structured and semantic latent spaces, which is our focus. Unlike traditional structured VAEs, we do not impose a specific structure *a priori*.

Our starting point is the following incongruence: On the one hand, it is well-established that the latent spaces of VAEs are typically entangled and semantically misaligned, fail to generalize OOD, and suffer from posterior collapse (which has been tied to latent variable nonidentifiability, 75). On the other hand, VAEs still learn highly structured latent spaces that can be traversed and interpolated, are "nearly" identifiable [80, 56], and represent abstract concepts linearly. Thus, our hypothesis is that sufficiently flexible models do capture meaningful structure internally, just not in a way that is interpretable or meaningful in practice. So, to learn latent structure, we build off the already performant embeddings of a generative model.

Our set-up is the following: We are given a black-box VAE, consisting of an encoder and a decoder, on which we make no assumptions other than the encoder outputs a latent code corresponding to $\mathbf{x}$. In the notation of (1), this corresponds to $\mathbf{z}$, however, to distinguish the black-box embeddings from our model, we will denote the black-box embeddings hereafter by $\mathbf{e}$. Our plan is to work solely with the embeddings $\mathbf{e}$ and learn how to extract linear concept representations from $\mathbf{e}$ such that distinct concepts can be intervened upon and composed together to create novel, OOD samples. Crucially, we do not modify the black-box architecture in any way. While this can be fine-tuned or trained end-to-end, it can also be frozen in place while our module trains separately.

3 RELATED WORK

Structured generative models To provide structure such as hierarchical, graphical, causal, and disentangled structures as well as other inductive biases in the latent space, there has been a trend towards building *structured* generative models that directly impose this structure *a priori*. Early work looked at incorporating fixed, known structure into generative models, such as autoregressive, graphical, and hierarchical structure [20, 26, 64, 77, 79, 15, 49]. This was later translated into known *causal* structure [31]. When the latent structure is unknown, several techniques have been developed to learn useful (not necessarily causal) structure from data [36, 24, 78, 30, 46]. More recently, based on growing interest in disentangled [7] and/or causal [59] representation learning, methods that learn causal structure have been developed [45, 83, 5, 63, 27].

Causal representation learning Causal representation learning [60], which involves learning causal structure in the latent space, is a rapidly developing area that has produced fundamental results on theoretical aspects of identifiability [9, 63, 32, 46, 30, 10, 22, 1, 74, 47, 38, 40, 84, 72, 65, 81, 86, 37, 73, 8, 2, 67, 85]. Recent work has also pushed in the direction of identifying concepts [33, 55, 19]. Our work draws inspiration from this line of work, which articulates precise conditions under which latent structure can be recovered *in principle*. By contrast, our focus is on methodological aspects of exploiting learned causal invariances *in practice*, although we do prove a new identifiability result that may be of independent interest.

Linear representations Generative models are known to represent concepts linearly in embedding space [e.g. 41, 66, 4]; see also [34, 4, 21, 3, 18, 62]. This phenomenon has been well-documented over the past decade in both language models [41, 52, 4, 14, 68, 16, 11, 69, 50, 48, 35, 51, 23, 25] and computer vision [53, 54, 6, 17, 28, 13, 76, 70]. Our approach actively exploits this tendency by searching for concept representations as linear projections of the embeddings learned by a black-box model.

OOD generalization A growing line of work studies the OOD generalization capabilities of generative models, with the general observation being that existing methods struggle to generalize OOD [82, 42, 43, 44, 61]. It is worth noting that most if not all of this work evaluates OOD generalization using reconstruction on held-out OOD samples, as opposed to generation. For example, a traditional VAE may be able to reconstruct held-out samples, but it is not possible to actively sample OOD. See Section 5.1 for more discussion.

4 ARCHITECTURE

We start with a black-box encoder-decoder architecture and assume there are d_c concepts of interest, denoted by $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_{d_c})$. Our objectives are two-fold:

1. To learn structure between concepts from black-box embeddings $\mathbf{e}$ as linear projections $C\mathbf{e}$;
2. To compose concepts together in a single, transparent model that captures how different concepts are related.

We seek to learn these concepts in an end-to-end, differentiable manner.

A key intuition is that composition can be interpreted as a type of intervention in the latent space. This is sensible since interventions in a causal model are a type of nontrivial distribution shift. Thus, the problem takes on a causal flavor which we exploit to build our architecture. The difficulty with this from the causal modeling perspective is that encoding structural assignments and/or causal mechanisms directly into a feed-forward neural network is tricky, because edges between nodes within the same layer aren't allowed in a feed-forward network but are required for the usual DAG representation of a causal model. To bypass this, we use the *reduced form* of a causal model which has a clear representation as a bipartite directed graph, with arrows only from exogenous to endogenous variables, making it conducive to being embedded into a feedforward NN. The tradeoff is that we do not learn a causal structure (i.e., a causal DAG), however, this enables the model to perform interventions directly in the latent space and to exploit invariances between interventional contexts.

4.1 OVERVIEW

Before outlining the architectural details, we provide a high-level overview of the main idea. A traditional decoder transforms the embeddings $\mathbf{e}$ into the observed variables $\mathbf{x}$, and the encoder operates in reverse by encoding $\mathbf{x}$ into $\mathbf{e}$. Thus,

$$\mathbf{x} \xrightarrow{\text{encoder}} \mathbf{e} \xrightarrow{\text{decoder}} \widehat{\mathbf{x}}.$$

Consistent with existing empirical work [41, 66, 4], we represent concepts as linear projections of the embeddings $\mathbf{e}$: Each concept $\mathbf{c}_j$ can be approximated as $\mathbf{c}_j \approx C_j\mathbf{e}$. This is modeled via a linear layer $\mathbf{c} \to \mathbf{e}$ that implicitly inverts this relationship in the decoder.

As a result, it makes sense to model the relationships between concepts with a linear SEM:

$$\mathbf{c}_j = \sum_{k=1}^{d_c} \alpha_{kj}\mathbf{c}_k + \varepsilon_k, \quad \alpha_{kj} \in \mathbb{R}.$$

By reducing this SEM, we deduce that

$$\mathbf{c} = A_0\boldsymbol{\varepsilon}, \text{ where } \begin{cases} \mathbf{c} = (\mathbf{c}_1, \dots, \mathbf{c}_{d_c}) \\ \boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_{d_c}) \end{cases}. \tag{2}$$

We model this with a linear layer $\boldsymbol{\varepsilon} \to \mathbf{c}$, where the weights in this layer correspond to the matrix A_0. This layer will be used to encode the SEM between the concepts, which will be used to implement causal interventions.

Remark 4.1. Due to the reduced form SEM above, our approach *does not* and *cannot* model the structural causal model encoded by the α_{kj}. What is important is that the reduced form $\mathbf{c} = A_0\boldsymbol{\varepsilon}$ still encodes causal invariances and interventions, which is enough in our setting, without directly estimating a causal graph.

In principle, since the exogenous variables ε_j are independent, we could treat $\boldsymbol{\varepsilon}$ as the input latent space to the generative model. Doing this, however, incurs two costs: 1) To conform to standard practice, $\boldsymbol{\varepsilon}$ would have to follow an isotropic Gaussian prior, and 2) It enforces artificial constraints on the latent dimension $\dim(\boldsymbol{\varepsilon})$. For this reason, we use a second expressive layer $\mathbf{z} \to \boldsymbol{\varepsilon}$ that gradually transforms $\mathbf{z} \sim \mathcal{N}(0, I)$ into $\boldsymbol{\varepsilon}$. This allows for $\dim(\mathbf{z})$ to be larger and more expressive than d_c (in practice, we set $\dim(\mathbf{z})$ to be a multiple of $\dim(\boldsymbol{\varepsilon})$), and for $\boldsymbol{\varepsilon}$ to be potentially non-Gaussian.

The final decoder architecture can be decomposed, at a high-level, as follows:

$$\mathbf{z} \longrightarrow \boldsymbol{\varepsilon} \longrightarrow \mathbf{c} \longrightarrow \mathbf{e} \longrightarrow \mathbf{x}. \tag{3}$$

Implementation details for each of these layers can be found in the next section.

4.2 DETAILS

Our goal is to augment the decoder of a generative model (without modifying the internals of the existing decoder) in such a way to identify d_c latent concepts from a black-box decoder. Moreover, we seek to accomplish this without sacrificing expressivity: The input layer to the decoder as well as the output layer of the encoder should both be arbitrary. This not only allows for arbitrary flexibility in both the latent dimension as well as the structure of the black-box encoder-decoder and its latent space (for example, a hierarchical or U-net architecture), but also makes implementation significantly simpler: The learning of latent concepts is completely abstracted away from the encoding step.

To accomplish this, we assume given a black-box encoder-decoder pair, denoted by $\mathrm{enc}(\mathbf{x})$ and $\mathrm{dec}(\mathbf{z})$, respectively. We modify the decoder by appending a *context module* to the head of the decoder. This context module is divided into three layers: A *representation layer*, an *intervention layer*, and an *expressive layer*.

1. The first *representation layer*, as the name suggests, learns to represent concepts by implicitly inverting their linear representations $\mathbf{c}_j = C_j\mathbf{e}$ from the embeddings of the encoder. This is a linear layer between $\mathbf{c} \to \mathbf{e}$.
2. The *intervention layer* embeds these concept representations into an overall reduced form SEM as suggested by (2). Each concept corresponds to its own context where it has been intervened upon. A key part of this layer is how it can be used to learn and enforce interventional semantics through this shared SEM. This layer corresponds to the second layer $\boldsymbol{\varepsilon} \to \mathbf{c}$ in (3).
3. The *expressive layer* is used to reduce the (potentially very large) input latent dimension of independent Gaussian inputs down to a smaller space of non-Gaussian exogenous noise variables for the intervention layer, corresponding to the first layer $\mathbf{z} \to \boldsymbol{\varepsilon}$ in (3). This is implemented as independent, deep MLPs that gradually

reduce the dimension in each layer: If it is desirable to preserve Gaussianity for the exogenous variables, linear activations can be used in place of nonlinear activations (e.g., ReLU).

Because the intervention layer is modeled after an SEM, it is straightforward to perform latent concept interventions using the calculus of interventions in an SEM.

Remark 4.2. Crucially, no part of this SEM is fixed or known—everything is trained end-to-end. In particular, we do not assume a known causal DAG or even a known causal order. This stands in contrast to previous work on causal generative models such as CausalGAN [31] and CausalVAE [83].

Concept interventions For simplicity, assume here that the concepts are each one-dimensional with $\dim(\mathbf{c}_j) = \dim(\boldsymbol{\varepsilon}_j) = 1$; generalization to multi-dimensional concepts is more or less straightforward and explained below. The structural coefficients $\alpha_{kj} \in \mathbb{R}$ capture direct causal effects between concepts, with $\alpha_{kj} \neq 0$ indicating the presence of an edge $\mathbf{c}_k \to \mathbf{c}_j$. An intervention on the jth concept requires deleting all these incoming edges; i.e., setting $\alpha_{\cdot j} = 0$, updating the outgoing edges from $\boldsymbol{\varepsilon}_j$, as well as replacing $\boldsymbol{\varepsilon}_j$ with a new $\boldsymbol{\varepsilon}'_j$, which results from updating the incoming edges to $\boldsymbol{\varepsilon}_j$ in the preceding layer. Thus, during training, when we intervene on $\mathbf{c}_j$, we zero out the row $\alpha_{\cdot j}$ while replacing the column $\alpha_{j\cdot}$ with a new column $\beta_{j\cdot}$ that is trained specifically for this interventional setting. The result is $d_c + 1$ intervention-specific layers: A_0 for the observational setting, and $A_1, \ldots, A_{d_c}$ for each interventional setting. This is captured as a three-way tensor where each slice of this tensor captures a different interventional context that corresponds to intervening on different concepts. At inference time, to generate interventional samples, we simply swap out the A_0 for A_j. Moreover, multi-target interventions can be sampled by zero-ing out multiple rows and substituting multiple $\beta_{j\cdot}$'s and $\boldsymbol{\varepsilon}_j$'s.

Multi-dimensional concepts In the causal model described above, we assumed $\dim(\mathbf{c}_j) = \dim(\boldsymbol{\varepsilon}_j) = 1$ for simplicity. Everything carries through if we allow additional flexibility with $\dim(\mathbf{c}_j) > 1$ and $\dim(\boldsymbol{\varepsilon}_j) > 1$, potentially even with $\dim(\mathbf{c}_j) \neq \dim(\boldsymbol{\varepsilon}_j)$. In practice, we implement this via two width parameters $w_\varepsilon = \dim(\boldsymbol{\varepsilon}_j)$ and $w_c = \dim(\mathbf{c}_j)$. The design choice of enforcing uniform dimensionality is not necessary, but is made here since in our experiments there did not seem to be substantial advantages to choosing nonuniform widths.

With these modifications, α_{kj} becomes a $w_\varepsilon \times w_c$ matrix; instead of substituting rows and columns above, we now substitute slices in the obvious way. The three-way tensor A is now $(d_c + 1) \times w_\varepsilon d_c \times w_c d_c$-dimensional.

Identifiability An appealing aspect of this architecture is that under certain assumptions, it can be viewed as an approximation to an identifiable model over disentangled concepts. Formally, we assume given observations $\mathbf{x}$, concepts $\mathbf{c}$, and a collection of embeddings $\mathbf{e}$ such that $\mathbf{x} = f(\mathbf{e})$. Additionally assume a noisy version of the linear representation hypothesis, i.e., $\mathbf{c}_j = C_j\mathbf{e} + \varepsilon_j$ with $\varepsilon_j \sim \mathcal{N}(0, \Omega_j)$. Then we have the following identifiability result:

Theorem 4.1 (Identifiability of linear representations)**.** *Assume that the rows of each C_j are chosen from a linearly independent set and f is injective and differentiable. Then, given single-node interventions on each concept $\mathbf{c}_j$, we can identify the representations C_j and the latent concept distribution $p(\mathbf{c})$.*

We do not claim to identify a causal graph, only the latent distribution $p(\mathbf{c})$ over the concepts and its functional structure C_j within the data-generating process. As opposed to solving a causal representation learning problem, the causal semantics used in this paper are simply being ported into the generative modeling framework to give solutions to the problem of learning latent *distribution* structure in a completely unsupervised fashion. Theorem 4.1 shows that the set problems where there *exists* a structured latent distribution to be identified is nonempty. Moreover, the structure of these identifiable distributions exactly fits our model architecture, meaning that set of problems in which the proposed method will approximate a fully disentangled latent space is nonempty. In other words, there *exist* data-modeling problems that are solvable by the novel approach introduced in this paper. The technical difficulty in the proof of Theorem 4.1 is analyzing the behavior of interventions in this model; since our main focus is on methodological aspects and the technical details are fairly involved, these details are deferred to the full version of the paper.

4.3 SUMMARY OF ARCHITECTURE

The context module thus proposed offers the following appealing desiderata in practice:

1. There is no coupling between the embedding dimension $\dim(\mathbf{e})$ and the number of concepts, or the complexity of the SEM. In particular, this architecture allows for *arbitrary* black-box encoder-decoder architectures to be used to learn embeddings, from which a causal model is then trained on top of. This can be trained end-to-end, fine-tuned after pre-training, or even frozen, allowing for substantial computational savings. In our experiments, we illustrate this both standard convolutional networks as well as NVAE [71], which is a deep hierarchical VAE with millions of parameters.
2. The intervention layer is a genuine causal model that provides rigourous causal semantics that allow for sam-

pling from arbitrary concept interventions, including interventions that have not been seen during training. But note that we do not claim to learn the structural coefficients, only the reduced model which is sufficient for concept interventions and sampling.

3. The architecture is based on an approximation to an identifiable concept model (Theorem 4.1), which provides formal justification for the intervention layer as well as reproducibility assurances.

4. The context module is itself arbitrarily flexible, meaning that there is no risk of information loss in representing concepts with the embeddings $\mathbf{e}$. Of course, information loss is possible if we compress this layer too much (e.g., by choosing d_c, k, w_ε, or w_c too small), but this is a design choice and not a constraint of the architecture itself.

As a consequence, the only tradeoff between representational capacity and causal semantics is design-based: The architecture itself imposes no such constraints. The causal model can be arbitrarily flexible and chosen independent of the black-box encoder-decoder pair, which is also allowed to be arbitrarily flexible.

5 EMPIRICAL RESULTS

We divide the presentation here into two main parts: small models on (semi-)synthetic datasets (Section 5.1) versus larger models on more challenging datasets (Section 5.2). We emphasize the key findings here but include the complete results and experimental details in the full version of the paper.

5.1 SMALL MODELS ON (SEMI-)SYNTHETIC DATA

To systematically investigate the interplay between architectural design, learned representations, and OOD performance, we begin with small-scale experiments with controlled environments. This setting enables rapid iteration, comprehensive ablation studies, and access to ground-truth metrics unavailable in complex real-world data. We evaluate models on two benchmarks:

- MNIST: We use Morpho-MNIST [12] to allow for different interventional contexts, such as making digits thicker or thinner.

- `quad`: A novel semi-synthetic visual environment with disentangled latents (color, shape, size, orientation) and explicit OOD compositionality challenges (e.g., generating unseen combinations of background quadrant colors and central object attributes).

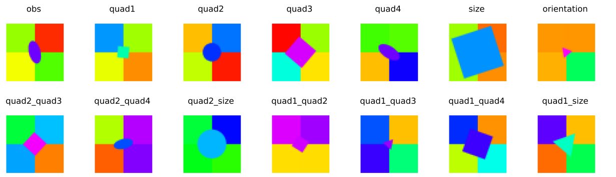


Figure 2: Example images obtained from the `quad` dataset. (top) Each subfigure corresponds to a different context, indicated by the title (`obs` = observational; the others indicate a single-node intervention). (bottom) Each subfigure corresponds to a different double-concept intervention context, indicated by the title. Double-concept interventions were not included in the training data, allowing for genuinely OOD evaluation.

MNIST On MNIST, we see (Table 1) results of our module across different ablations. The first column (Context module) is our module incorporated into a lightweight convolutional VAE. Ablation 1 is an analogously lightweight standard implementation of β-VAE trained only on the observational context. Ablation 2 is β-VAE trained on the full pooled dataset. Ablation 3 is similar to the context module, except that it ignores the contexts and pools everything into a single context (like Ablation 2). These results indicate that the cost of the structured representation learned by the context module is only a slight degradation in the numerical metrics.

Controlled simulations on `quad` While evaluations on unsimulated data are the gold standard for evaluating models, simulated environments are valuable for conducting controlled experiments under given conditions. To provide a controllable, synthetic test bed for evaluating composition in a visual environment, we developed a simple semi-synthetic benchmark, visualized in Figure 2. `quad` is a visual environment defined by 8 concepts:

1. `quad1`: The color of the first quadrant;
2. `quad2`: The color of the second quadrant;
3. `quad3`: The color of the third quadrant;
4. `quad4`: The color of the fourth quadrant;
5. `size`: The size of the center object;
6. `orientation`: The orientation (angle) of the center object;
7. `object`: The color of the center object;
8. `shape`: The shape of the center object (circle, square, pill, triangle).

With the exception of `shape`, which is discrete, the remaining concepts take values in $[0, 1]$.

Table 2 shows numerical results comparing our module (incorporated into a lightweight convolutional VAE) to Ablation 2 from above (analogously lightweight β-VAE trained

Table 1: Ablation on MNIST (200 epochs)

Metric / Evaluation	Context module	Ablation 1	Ablation 2	Ablation 3
Validation Metrics				
Validation ELBO (bpd) ↓	0.2945 (0.0091)	0.1832 (0.0007)	0.2362 (0.0053)	0.2886 (0.0103)
Reconstruction Loss (bpd) ↓	0.2649 (0.0093)	0.1469 (0.0008)	0.1998 (0.0050)	0.2575 (0.0111)
Reconstructed Distribution Evaluation				
MMD (obs) ↓	0.0092 (0.0027)	0.0000 (0.0000)	0.0028 (0.0022)	0.0124 (0.0032)
MMD (ivn) ↓	0.0939 (0.0331)	0.0303 (0.0107)	0.0430 (0.0145)	0.0536 (0.0154)
Generated Distribution Evaluation				
MMD (obs) ↓	0.0099 (0.0011)	0.0017 (0.0007)	0.0097 (0.0030)	0.0098 (0.0019)
MMD (ivn) ↓	0.0298 (0.0069)	0.0142 (0.0038)	0.0183 (0.0034)	0.0188 (0.0034)

Table 2: Ablation on Quad

Metric / Evaluation	Context module	Ablation 2
Generated Sample Evaluation		
MMD (obs) ↓	0.0099 (0.0058)	0.0187 (0.0063)
MMD (ivn) ↓	0.0149 (0.0032)	0.0275 (0.0038)
MMD (ood) ↓	0.0273 (0.0021)	0.0297 (0.0019)

on the full pooled dataset) for sample generation across the observational and interventional data (same distribution seen during training) along with OOD data. The `quad` data allows evaluating the sampled OOD distribution (which our module composes from seen contexts) with a ground truth OOD distribution (unseen during training).

OOD generation vs. reconstruction There is a crucial difference between OOD reconstruction and OOD generation. By OOD generation, we mean the ability to conditionally generate and combine novel interventions. For example, during training the model may have only seen small, red objects along with big, blue objects. At test time, we wish to generate OOD samples of big, red objects or small, blue objects.

The crucial difference is that a model may be capable of OOD reconstruction, but not OOD generation. For example, we can always attempt to reconstruct an OOD sample and evaluate its reconstruction error. This is a common metric that has been used in previous work to evaluate the OOD capabilities of generative models [e.g. 82, 42, 43, 61]. But unless the model learns specific latent concepts corresponding to size or colour, then we have no control over the size or colour of random samples from the model. We can always condition on the representations learned by the model, but these representations will not be disentangled or interpretable. Thus, OOD generation not only evaluates the ability of a model to compose learned concepts in new ways, but also implicitly evaluates the ability of a model to identify and capture underlying concepts of interest. For this reason, we argue that OOD generation is a more appropriate evaluation metric.

While the numerical results on synthetic data above indicate better performance on OOD generation using our module, a more intuitive visual evaluation on real data is provided in Section 5.2, Figure 4.

5.2 LARGE MODELS ON REAL DATA

We now evaluate performance when our module is incorporated into a complex black-box model (NVAE of Vahdat and Kautz [71]) on more challenging datasets, again including MNIST but now also 3DIdent [87] and CelebA [39]. Since CelebA does not have interventional data, we use this as an important ablation to test how well our module performs when used with *conditional* data (i.e. by conditioning on attributes in CelebA).

Table 3: Comparison of models (BPD = bits per dimension)

Dataset	Type	Black-box	BPD
MNIST	Observational	NVAE	0.144
MNIST	Interventional	NVAE	0.149
3DIdent	Observational	NVAE	0.673
3DIdent	Interventional	NVAE	0.754
CelebA	Observational	NVAE	2.08
CelebA	Conditional	NVAE	2.13

Table 3 summarizes the results of these training runs, where “observational” means that the standard NVAE (i.e. without a context module) is trained on the entire datasets, whereas “interventional” means that the context module was used.

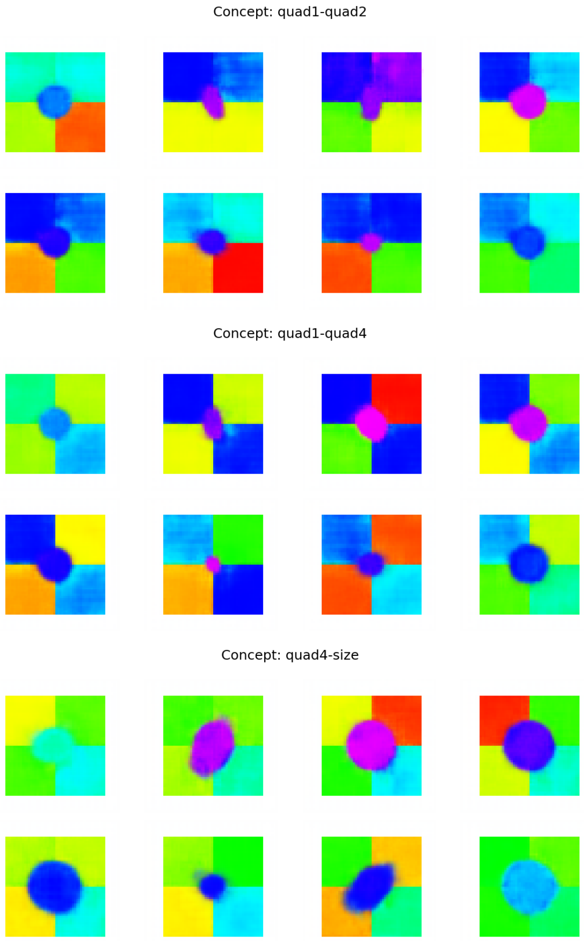


Figure 3: Example generated OOD images from a simple, lightweight VAE on the `quad` benchmark. This shows a proof-of-concept that the concept module successfully composes concepts using even simple architectures. For performant architectures, see the NVAE results in Figure 4.

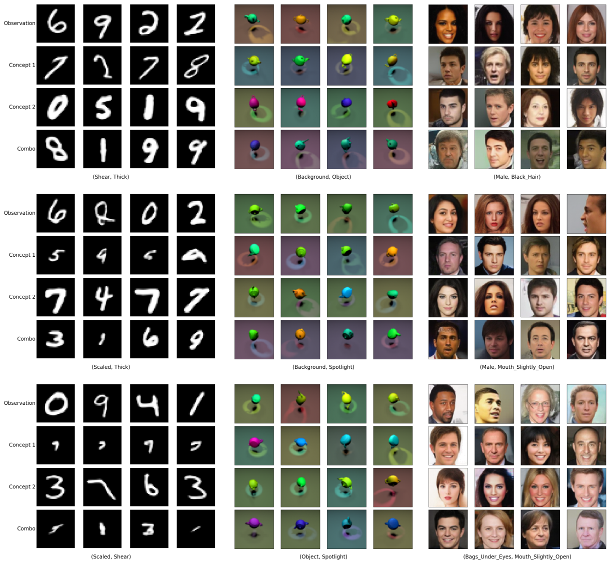


Figure 4: Additional examples of concept composition in MNIST (left), 3DIdent (middle), and CelebA (right). In MNIST and 3DIdent, the samples are OOD: The training data did not contain any examples with these concepts composed together. The CelebA results are an ablation to understand the effect of conditioning vs intervention, and so there is some leakage between concepts, where the interventional datasets (MNIST, CelebA) show no leakage. (Due to space constraints, a larger version of this figure can be found in Figure 5 in the appendix.)

We observe competitive reconstruction, with only slight degradation when the context module is attached to the NVAE decoder.

More importantly, Figure 4 demonstrates the generation capabilities enabled by our module. Each row corresponds to actual generated samples from the trained model in different contexts:

1. The first row shows generated observational samples;
2. The second and third row contain generated single-node interventions;
3. The final row shows generated samples where two distinct concepts are composed together.

For MNIST and 3DIdent, the final row of compositional generations is genuinely OOD in that the training data does not contain any examples where both concepts have been intervened upon. CelebA is trained using conditional as opposed to interventional samples as an ablation on the sensitivity to non-interventional data. Although there is a slight drop in OOD perceptual accuracy on CelebA, this is to be expected since the model was trained on conditional data as opposed to interventions. Finally, observe the last row of Figure 4, which demonstrates the OOD generation capability enabled by our module: these contexts did not appear in the training data but our module is able to sensibly generate them by composing and generalizing from the concepts it learned according to different data contexts.

Acknowledgements

The work by AM was supported by Novo Nordisk Foundation Grant NNF20OC0062897.

This research was supported in part through the computational resources and staff contributions provided for the Mercury high performance computing cluster at The University of Chicago Booth School of Business which is supported by the Office of the Dean.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR, 2023.

[2] Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2024.

[3] Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR, 2019.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[5] Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural admg learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[7] Yoshua Bengio. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing: First International Conference, SLSP 2013, Tarragona, Spain, July 29-31, 2013. Proceedings 1*, pages 1–37. Springer, 2013.

[8] Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Causal Learning and Reasoning*, pages 843–867. PMLR, 2024.

[9] 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.

[10] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

[11] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[12] Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20 (178):1–29, 2019.

[13] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[14] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[15] Mucong Ding, Constantinos Daskalakis, and Soheil Feizi. GANs with conditional independence graphs: On subadditivity of probability divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 3709–3717. PMLR, 2021.

[16] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

[17] Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. *arXiv preprint arXiv:1711.05772*, 2017.

[18] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*, 2018.

[19] Hidde Fokkema, Tim van Erven, and Sara Magliacane. Sample-efficient learning of concepts with theoretical guarantees: from data to concepts without interventions. 02 2025. URL https://arxiv.org/pdf/2502.06536.pdf.

[20] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.

[21] Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. Skip-gram- zipf+ uniform= vector additivity.

In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76, 2017.

[22] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

[24] Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019.

[25] Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. *arXiv preprint*, 2024.

[26] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29, 2016.

[27] David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15639–15654. PMLR, 23–29 Jul 2023.

[28] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

[30] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

[31] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv:1709.02023 [cs.LG]*, 2017.

[32] Sébastien Lachapelle, Pau Rodríguez, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *1st Conference on Causal Learning and Reasoning, CLeaR 2022*, volume 177 of *Proceedings of Machine Learning Research*, pages 428–484. PMLR, 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

[33] Tobias Leemann, Michael Kirchhof, Yao Rong, Enkelejda Kasneci, and Gjergji Kasneci. When are post-hoc conceptual explanations identifiable? In *Uncertainty in Artificial Intelligence*, pages 1207–1218. PMLR, 2023.

[34] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

[35] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

[36] Xiaopeng Li, Zhourong Chen, Leonard KM Poon, and Nevin L Zhang. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. *arXiv preprint arXiv:1803.05206*, 2018.

[37] Xiu-Chuan Li, Kun Zhang, and Tongliang Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *The Twelfth International Conference on Learning Representations*, 2024.

[38] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. *arXiv preprint arXiv:2310.15580*, 2023.

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[40] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[42] Zhoutong Mo et al. Compositional generative modeling: A single model is not all you need, 2024. URL https://arxiv.org/abs/2402.01103.

[43] Milton Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, 35:10136–10149, 2022.

[44] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021.

[45] Raha Moraffah, Bahman Moraffah, Mansooreh Karami, Adrienne Raglin, and Huan Liu. Causal adversarial network for learning conditional and interventional distributions. *arXiv preprint arXiv:2008.11376*, 2020.

[46] Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

[47] Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023.

[48] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodola. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

[49] Jacobie Mouton and Rodney Stephen Kroon. Integrating Bayesian network structure into residual flows and variational autoencoders. *Transactions on Machine Learning Research*, 2023.

[50] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

[51] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[53] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[54] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep understanding and improvement. *stat*, 1050:19, 2017.

[55] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

[56] Patrik Reizinger, Luigi Gresele, Jack Brady, Julius von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: Vaes perform independent mechanism analysis. *arXiv preprint arXiv:2206.02416*, 2022.

[57] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[58] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[59] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[60] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. arXiv:2102.11107.

[61] Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *10th International Conference on Learning Representations (ICLR 2022)*. OpenReview. net, 2022.

[62] Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word vectors. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 387–396, 2019.

[63] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55, 2022.

[64] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

[65] Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Advances in Neural Information Processing Systems*, 36:34465–34492, 2023.

[66] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[67] Davide Talon, Phillip Lippe, Stuart James, Alessio Del Bue, and Sara Magliacane. Towards the reusability and compositionality of causal representations. *arXiv preprint arXiv:2403.09830*, 2024.

[68] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[69] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

[70] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2023.

[71] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[72] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

[73] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

[74] Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

[75] Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34:5443–5455, 2021.

[76] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for score-based conditional model. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

[77] Stefan Webb, Adam Golinski, Rob Zinkov, Tom Rainforth, Yee Whye Teh, Frank Wood, et al. Faithful inversion of generative models for effective amortized inference. *Advances in Neural Information Processing Systems*, 31, 2018.

[78] Antoine Wehenkel and Gilles Louppe. Graphical normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 37–45. PMLR, 2021.

[79] Christian Weilbach, Boyan Beronov, Frank Wood, and William Harvey. Structured conditional continuous normalizing flows for efficient amortized inference in graphical models. In *International Conference on Artificial Intelligence and Statistics*, pages 4441–4451. PMLR, 2020.

[80] Matthew Willetts and Brooks Paige. I don't need $\mathbf{u}$: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

[81] Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024.

[82] Zhenlin Xu, Marc Niethammer, and Colin Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language, 2022. URL https://arxiv.org/abs/2210.00482.

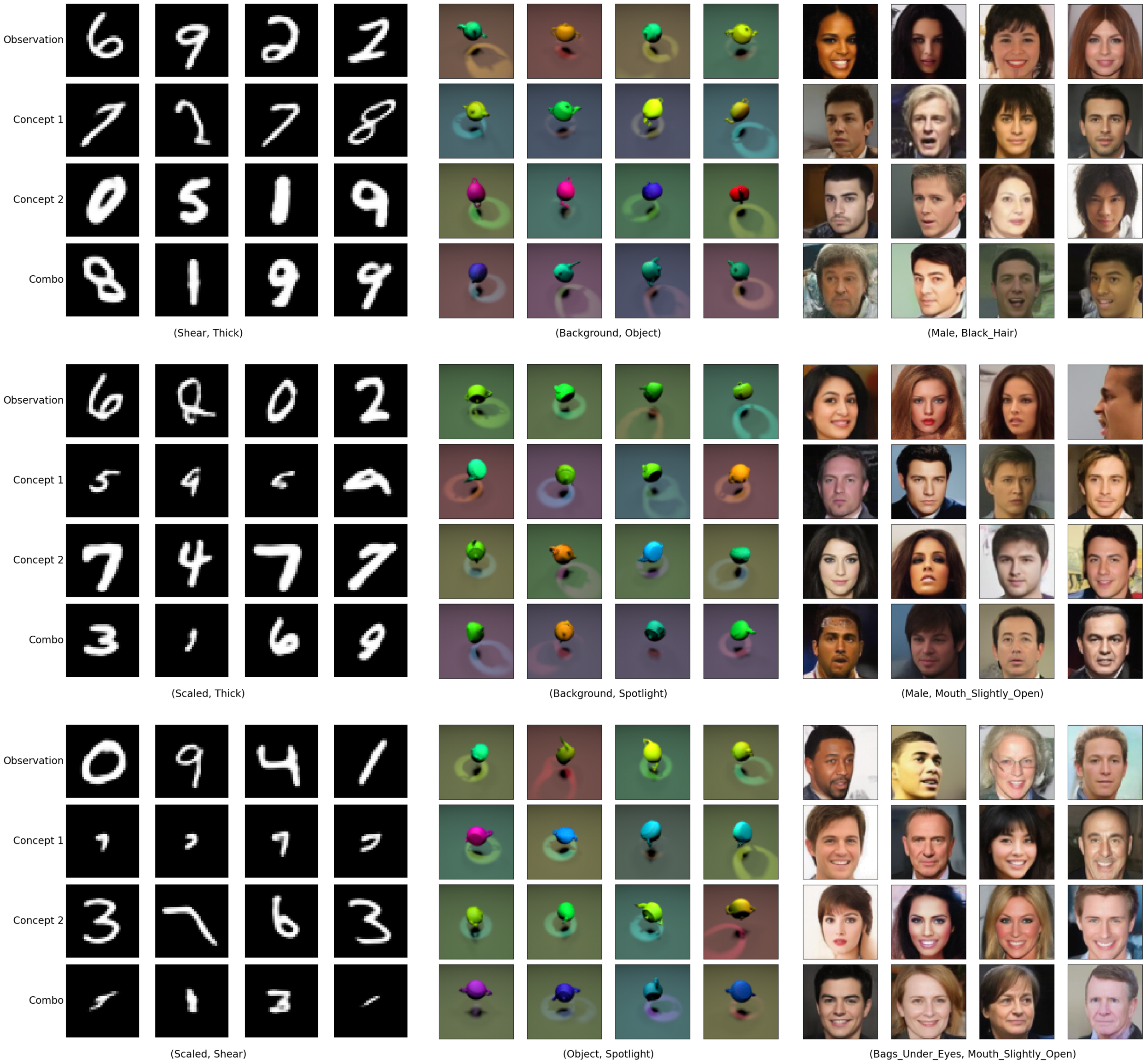
[83] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[84] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

[85] Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024.

[86] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

[87] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pages 12979–12990. PMLR, 2021.



(Shear, Thick)

(Background, Object)

(Male, Black_Hair)

(Scaled, Thick)

(Background, Spotlight)

(Male, Mouth_Slightly_Open)

(Scaled, Shear)

(Object, Spotlight)

(Bags_Under_Eyes, Mouth_Slightly_Open)

Figure 5: Full size version of Figure 4; caption reproduced here: Additional examples of concept composition in MNIST (left), 3DIdent (middle), and CelebA (right). In MNIST and 3DIdent, the samples are OOD: The training data did not contain any examples with these concepts composed together. The CelebA results are an ablation to understand the effect of conditioning vs intervention, and so there is some leakage between concepts, where the interventional datasets (MNIST, CelebA) show no leakage.