Concept Algebra for Score-based Conditional Model

Zihao Wang [1] **Lin Gui** [1] **Jeffrey Negrea** [2] **Victor Vietch** [1 2 3]

Abstract

This paper concerns the structure of learned representations in text-guided generative models, focusing on score-based models. A key property of such models is that they can compose disparate concepts in a ‘disentangled’ manner. This suggests these models have internal representations that encode concepts in a ‘disentangled’ manner. Here, we focus on the idea that concepts are encoded as subspaces of some representation space. We formalize this, show there’s a natural choice for the representation, and develop a simple method for identifying the part of the representation corresponding to a given concept. In particular, this allows us to manipulate the concepts expressed by the model through algebraic manipulation of the representation. We demonstrate it with examples using Stable Diffusion.

1. Introduction

Large-scale text-controlled generative models are now dominant in many parts of modern machine learning and artificial intelligence (e.g., Brown et al., 2020; Radford et al., 2021; Bommasani et al., 2021; Kojima et al., 2022). In these models, the user provides a prompt in natural language and the model generates samples based on this prompt—e.g., in large language models the sample is a natural language response, and in text-to-image models the sample is an image. These models have a remarkable ability to compose disparate concepts to generate coherent samples that were not seen during training.This suggests that these models have some internal representation of high-level concepts that can be manipulated in a ‘disentangled’ manner. Broadly, the goal of this paper is to shed light on how this concept representation works, and how it can be manipulated. We focus on text-to-image diffusion models, though many of the ideas are generally applicable.

Our starting point is the following commonly observed structure of representations:

1. Each data point x is mapped to some representation vector $\mathrm{Rep}(x) \in \mathbb{R}^p$.
2. High-level concepts correspond to subspaces (directions) of the representation space.

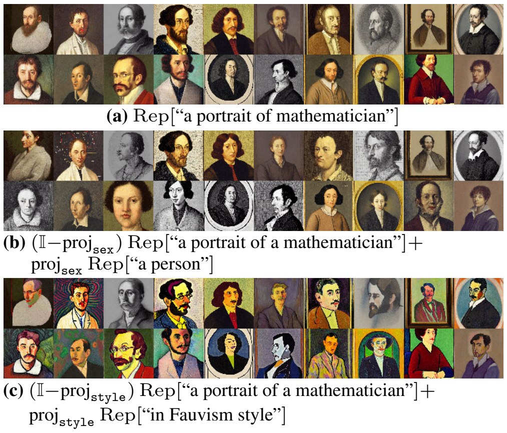
Perhaps the best known example of this structure is in word embeddings, where semantic relationships such as $\mathrm{Rep}(\text{“king”}) - \mathrm{Rep}(\text{“queen”}) \approx \mathrm{Rep}(\text{“man”}) - \mathrm{Rep}(\text{“woman”})$ suggest that high-level concepts (here, sex) are encoded as directions in the representation space (Mikolov et al., 2013a). This kind of encoding of concepts has been argued to occur in many contexts, including in the latent space of variational autoencoders (Zhou & Wei, 2020; Khemakhem et al., 2020; Mita et al., 2021) and in the latent space of language models (Bolukbasi et al., 2016; Gonen & Goldberg, 2019; Radford et al., 2021; Elhage et al., 2022). We’ll call representations of this kind *arithmetically composable*, because composition corresponds to arithmetic operations on the representation vectors. The goal of this paper is to develop arithmetically composable representations of text for score-based text to image models.

There are two main motivations. First, understanding the structure of the representation space is important for foundational progress on understanding the emergent behavior of text-controlled generative models. It is particularly interesting to study this question in the text-to-image setting because the multi-modality of the data makes it straightforward to distinguish concepts from inputs, and because it is not clear a priori that the models themselves build in any inductive bias towards arithmetic structure. The secondary motivation is that having such a representation would allow us to manipulate the concepts expressed by the model through linear-algebraic manipulations of representation of the input text; Figure 1 illustrates this idea.

The development of the paper is as follows: first, we develop a mathematical formalism for describing the connection between representation structures and concepts for text-controlled generative models. Then, using this formalism, we show that the Stein score of the text-conditional distribution is an arithmetically composable representation of the input text. Finally, we develop *concept algebra* as

[1] Department of Statistics, University of Chicago [2] Data Science Institute, University of Chicago [3] Google Brain. Correspondence to: Zihao Wang <wangzh@statistics.uchicago.edu>, Victor Vietch <victorveitch@gmail.com>.

Accepted to ICML workshop on Spurious correlations, Invariance, and Stability, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).


(a) Rep["a portrait of mathematician"]

(b) $(\mathbb{I}-\text{proj}_{\texttt{sex}})$ Rep["a portrait of a mathematician"]+ $\text{proj}_{\texttt{sex}}$ Rep["a person"]

(c) $(\mathbb{I}-\text{proj}_{\texttt{style}})$ Rep["a portrait of a mathematician"]+ $\text{proj}_{\texttt{style}}$ Rep["in Fauvism style"]

Figure 1. We show that high-level concepts such as `sex` and `artistic_style` are encoded as subspaces of some representation space. This allows us to manipulate the concepts expressed by a prompt through algebraic operations on the representation of that prompt. Namely, we edit the representation projected on to the subspace corresponding to a concept. Note images are paired by random seed.

a method for manipulating the concepts expressed by the model through algebraic manipulation of this representation. We provide theoretical justifications (Appendix A) for this approach, and illustrate it with examples (Appendix B) manipulating a variety of concepts.

2. A Mathematical Framework for Concepts as Subspaces

Our first task involves creating a precise mathematical framework bridging representations and high-level concepts. It's vital to comprehend their feasibility, construction methods, and possible failures. We must accurately define a concept, its relation to inputs x, and its representation process.

Concepts The real-world process that generated the training data has the following structure. First, images Y are generated according to some real-world, physical process. Then, some human looks at each image and writes a caption describing it. Inverting this process, each text x induces some probability density $p(y \mid x)$ over images Y based on how compatible they are with x as a caption. The (implicit) goal of the generative model is to learn this distribution.

To write the caption, the human first maps the image to a set of high-level variables summarizing the image's content, then uses these latent variables to generate the text X. Let C be the latent variable that captures all the information about the image that is relevant for a human writing a caption. So,

$$p(y \mid X = x) = \int p(y \mid C = c)p(C = c \mid X = x)\mathrm{d}c.$$

The random variable C captures the information that is jointly relevant for both the image and caption. Variables in C include attributes such as `has_mathematician` or `is_man`, but not `pixel_14_is_red`. We define concepts in terms of the latent C.

Definition 2.1. A *concept variable* Z is a C-measurable random variable. The *concept* $\mathcal{Z}$ associated to Z is the sample space of Z.

Now, the full set of all possible concepts is unwieldy. Generally, we are concerned only with the concepts elicited by a particular prompt x.

Definition 2.2. A set of concepts $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$ is *sufficient for* x if $p(c \mid X = x, Z_{1:k} = z_{1:k}) = p(c \mid Z_{1:k} = z_{1:k})$ for all $z_1, \ldots, z_k \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_k$.

For example, the concept `profession` would be sufficient for the prompt "A nurse". This prompt induces a distribution on many concepts (e.g., `background` is likely to be a hospital) but these other concepts are independent of the caption given `profession` = `nurse`. Then,

$$p(y \mid x) = \sum_{z_{1:k}} p(y \mid z_{1:k})p(z_{1:k} \mid x). \tag{1}$$

Concept Distributions Following Equation (1), we can view each text x as specifying a distribution $p(z_{1:k} \mid x)$ over latent concepts $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$. This observation lets us make the relationship between text and concepts precise.

Definition 2.3. A *concept distribution* Q is a distribution over concepts. Each text x specifies a concept distribution as $Q_x = p(z_{1:k} \mid x)$.

That is, we move from viewing text as expressing specific concept values ($\texttt{is\_mathematician} = 1$) to expressing probability distributions over concepts ($Q_x(\texttt{is\_mathematician} = 1) = 0.99$). The probabilistic view is more general—deterministically expressed concepts can be represented as degenerate distributions. This extra generality is necessary: e.g., the prompt "a person" induces a non-degenerate distribution over the `sex` concept.

Concept Representations A text-controlled generative model takes in prompt text x and produces a random output Y. Implicitly, such models are maps from text strings x to the space of probability densities over Y. We'll define a representation $\text{Rep}(x) \in \mathcal{R}$ of x as any function of x that suffices to specify the output distribution. We define $f_r(\cdot)$ as the density defined by $r \in \mathcal{R}$, and assume that the model learn's the true data distribution of $Y|X = x$:

$$f_{\text{Rep}(x)}(y) = p(y \mid x). \tag{2}$$

The key idea for connecting representations and concepts is to move from considering representations of prompts to representations of concept distributions.

Definition 2.4. A *concept representation* Rep is a function that maps a concept distribution Q to a representation $\mathrm{Rep}(Q) \in \mathcal{R}$, where $\mathcal{R}$ is a vector space. The *representation of a prompt* x is the representation of the associated concept distribution, $\mathrm{Rep}(x) := \mathrm{Rep}(Q_x)$.

There are two reasons why this view is desirable. First, defining the representation in terms of the concept distribution makes the role of concepts explicit—this will allow us to explain how representation structure relates to concept structure. Second, it allows us to reason about representations that don't correspond to any prompt. Every prompt defines a concept distribution, but not the other way around. This matters because we ultimately want to reason about the conceptual meaning of representation vectors created by algebraic operations on representations of prompts. Such vectors need not correspond to any prompt.

Arithmetic Compositionality We now have the tools to define what it means for a representation to be arithmetically composable. We define composability for a pair of concepts $\mathcal{Z}$ and $\mathcal{W}$. In the subsequent development, our aim will be to manipulate $\mathcal{Z}$ while leaving $\mathcal{W}$ fixed.

Definition 2.5. A representation Rep is *arithmetically composable* with respect to concepts $\mathcal{Z}, \mathcal{W}$ if there are vector spaces $\mathcal{R}_Z$ and $\mathcal{R}_W$ such that for all concept distributions of the form $Q(z, w) = Q_Z(z)Q_W(w)$,

$$\mathrm{Rep}(Q_Z Q_W) = \mathrm{Rep}_Z(Q_Z) + \mathrm{Rep}_W(Q_W),$$

where $\mathrm{Rep}_Z(Q_Z) \in \mathcal{R}_Z$ and $\mathrm{Rep}_W(Q_W) \in \mathcal{R}_W$.

In words: we restrict to product distributions to capture the requirement that the concepts $\mathcal{Z}$ and $\mathcal{W}$ can be manipulated freely of each other (the typical case is that one or both of Q_Z and Q_W are degenerate, putting all their mass on a single point). Then, the definition requires that there are fixed subspaces corresponding to each concept in the sense that, e.g., changing only Q_Z induces a change only in $\mathcal{R}_Z$.

3. The Score Representation

We now have an abstract definition of arithmetically composable representation. The next step is to find a specific representation function that satisfies the definition.

We will study the following choice.

Definition 3.1. The *score representation* $s[Q]$ of a concept distribution Q is defined by:

$$s[Q](y) := \nabla_y \log \int p(y \mid z, w) Q(z, w) \mathrm{d}z \mathrm{d}w.$$

The *centered score representation* $\bar{s}[Q]$ is defined by $\bar{s}[Q] := s[Q] - s[Q_0]$.

Here, $s[Q]$ is itself a function of y and the representation space $\mathcal{R}$ is a vector space of functions. This is a departure from the typical view of representations as elements of $\mathbb{R}^p$. The score representation can be thought of as a kind of non-parametric representation vector. The centered score representation just subtracts off the representation of some baseline distribution Q_0.[1]

The main motivation for studying the score representation is that

$$s[x](y) := s[Q_x](y) = \nabla_y \log p(y \mid x).$$

The importance of this observation is that $\nabla_y \log p(y \mid x)$ is learnable from data. In fact, this score function is ultimately the basis of many controlled generation models (e.g., Ho et al., 2020; Ramesh et al., 2022; Saharia et al., 2022), because it characterizes the conditional while avoiding the need to compute the normalizing constant (Hyvärinen & Dayan, 2005; Song & Ermon, 2019). Accordingly, we can readily compute the score representation of prompts in many generative models, without any extra model training.

Causal Separability The score representation does not have arithmetically composable structure with respect to every pair of concepts. The crux of the issue is that concepts are reflected in the representation based on their effect on Y. If the way they affect Y depends fundamentally on some interaction between two concepts, the representation cannot hope to disentangle them. Thus, we must rule out this case.

Definition 3.2. We say that Y is *causally separable* with respect to $\mathcal{Z}, \mathcal{W}$ if there exist unique Y-measurable variables $Y_{\mathcal{Z}}$, $Y_{\mathcal{W}}$, and ξ such that

1. $Y = g(Y_{\mathcal{Z}}, Y_{\mathcal{W}}, \xi)$ for some invertible and differentiable function g, and
2. $p(y_{\mathcal{Z}}, y_{\mathcal{W}}, \xi \mid z, w) = p(\xi)p(y_{\mathcal{Z}} \mid z)p(y_{\mathcal{W}} \mid w)$

Informally, the requirement is that we can separately generate $Y_{\mathcal{Z}}$ and $Y_{\mathcal{W}}$ as the part of the output affected by $\mathcal{Z}$ and $\mathcal{W}$(and ξ as the part of the image unrelated to Z and W), then combine these parts to form the final image. That is, generating the visual features associated to a concept $\mathcal{W}$ can't require us to know the value of another concept $\mathcal{Z}$. As an example where causal separability fails, consider the concepts of `species` $\mathcal{W} = \{\texttt{deer}, \texttt{human}\}$ and `sex` $\mathcal{Z} = \{\texttt{male}, \texttt{female}\}$. It seems reasonable that there is a Y-measurable $Y_{\mathcal{W}}$ that is the species part of the image—e.g., the presence of fur vs skin, snouts vs noses, and so forth. However, there is no part of Y that corresponds to a sex concept in a manner that's free of species. The reason is that the visual characteristics of sex are fundamentally

[1] The representation space $\mathcal{R}$ is the same for all Q_0; the choice is arbitrary. We define $\bar{s}$ to ensure 0 is an element of $\mathcal{R}$. This is for theoretical convenience; we will see that only s is required in practice.

different across species—e.g., male deer have antlers, but humans usually do not. In Figure 5 we test this example, finding that concept algebra fails in the absence of causal separability.

It turns out it suffices to rule out this case (all proofs in appendix):

Proposition 3.1. *If Y is causally separable with respect to $\mathcal{W}$ and $\mathcal{Z}$, then the centered score representation is arithmetically composable with respect to $\mathcal{W}$ and $\mathcal{Z}$.*

That is: the (centered) score representation is structured such that concepts correspond to subspaces of the representation space.

4. Concept Algebra

We have established that concepts correspond to subspaces of the representation space. We now consider how to manipulate concepts through algebraic operations on representations.

To modify a particular concept $\mathcal{Z}$ we want to modify the representation only on the subspace $\mathcal{R}_Z$ corresponding to $\mathcal{Z}$. For example, consider changing the style concept to `Fauvism`. Intuitively, we want an operation of the form:

$$\begin{aligned} s_{\text{edit}} \leftarrow & (\mathbb{I} - \text{proj}_{\texttt{style}}) s[\text{"a portrait of mathematician"}] \\ & + \text{proj}_{\texttt{style}} s[\text{"Fauvism style"}], \end{aligned}$$

where $\text{proj}_{\texttt{style}}$ is the projection onto the subspace corresponding to the style concept. The idea is that the representation of the original prompt x_{orig} is unchanged except on the style subspace. On the style subspace, the representation takes on the value elicited by the new prompt $x_{\text{new}} =$ "Fauvism style".

There are two main challenges for putting this intuition into practice. First, because we are working with an infinite dimensional representation, it is unclear how to do the projection. Second, although we know that some $\mathcal{R}_Z$ exists, we still need a way to determine it explicitly.

4.1. Concept Manipulation through Projection

Following Proposition 3.1, we have that

$$\bar{s}[Q_Z \times Q_W] = \bar{s}_Z[Q_Z] + \bar{s}_W[Q_W], \tag{3}$$

for some representation functions $\bar{s}_Z$ and $\bar{s}_W$ with range in $\mathcal{R}_Z$ and $\mathcal{R}_W$ respectively. We have that the Z-representation space is

$$\mathcal{R}_Z = \text{span}(\{\bar{s}_Z[Q_Z] : Q_Z \text{ a distribution}\}). \tag{4}$$

Our goal is to find a projection onto $\mathcal{R}_Z$.

The first obstacle is that $\mathcal{R}_Z$ is a function space, making algebraic operations difficult to define. The resolution is straightforward. In practice, score-based models generate samples by running a discretized (stochastic) differential equation forward in time. These algorithms only require the score function evaluated at the finite set of points. At each y, we have that $\bar{s}(y) \in \mathbb{R}^m$. Accordingly, by restricting attention to a single value of y at a time, we can use ordinary linear algebra to define the manipulations:

Definition 4.1. The *Z subspace at y* is

$$\mathcal{R}_Z(y) := \text{span}(\{\bar{s}_Z[Q_Z](y) : Q_Z \text{ a distribution}\}) \tag{5}$$

and the *Z-projection at y*, denoted $\text{proj}_Z(y)$ is the projection onto this subspace.

If we can compute $\text{proj}_Z(y)$ then we can just edit the representation at each point y. That is, we transform the score function at each point:

$$\bar{s}_{\text{edit}}(y) \leftarrow (\mathbb{I} - \text{proj}_Z(y)) \bar{s}[x_{\text{orig}}](y) + \text{proj}_Z(y) \bar{s}[x_{\text{new}}](y). \tag{6}$$

We then draw samples from the stochastic differential equation defined by $\bar{s}_{\text{edit}}$.

4.2. Concept Algebra

Now the challenge is to identify $\mathcal{R}_Z(y)$. The idea is to find a basis for the subspace using prompts $x_0, \dots x_k$ that elicit distributions of the form $Q_{x_j} = Q_Z^j Q_W$. For example, to identify the `sex` concept we use the prompts $x_0 =$ "a man" and $x_1 =$ "a woman", with the idea that

$$Q_{x_0} = \delta_{\text{male}} \times Q_W, \;\; Q_{x_1} = \delta_{\text{female}} \times Q_W,$$

with the same marginal distribution Q_W. We then use the prompts to define the estimated subspace as

$$\hat{\mathcal{R}}_Z(y) := \text{span}(\{s[x_i](y) - s[x_0](y) : i = 1, \dots, k\}).$$

Summarizing, our approach to algebraically manipulating concepts is:

1. Find prompts $x_0, \dots, x_k$ such that each elicits a different distribution on Z, but the same distribution on W. That is, $Q_{x_j} = Q_Z^j Q_W$ for each j.
2. Construct the estimated representation space $\hat{\mathcal{R}}_Z(y)$ following Section 4.2, and define $\text{proj}_Z(y)$ as the projection onto this space.
3. Sample from the discretized SDE defined by the manipulated score representation[2]

$$s_{\text{edit}}(y) \leftarrow (\mathbb{I} - \text{proj}_Z(y)) s[x_{\text{orig}}](y) + \text{proj}_Z(y) s[x_{\text{new}}](y). \tag{7}$$

Appendix A provides theoretical justification for Equation (7), and Appendix C describes how to implement it.

[2] We can view this as first editing the centered representation $\bar{s}$: $\bar{s}_{\text{edit}}(y) \leftarrow (\mathbb{I} - \text{proj}_Z(y)) \bar{s}[x_0](y) + \text{proj}_Z(y) \bar{s}[\tilde{x}](y)$. Then add the same baseline on both sides.

5. Discussion and Related Work

We have introduced a mathematical framework to make precise the notion that concepts correspond to subspaces of a representation space. Using this framework, we proved that the score representation has this structure, and derived a method for determining the subspace corresponding to a given concept (Appendix A). Finally, we showed how to use this structure to manipulate expressed concepts in the score representation of a diffusion model.

Concepts as Subspaces There has been significant interest in whether and how neural representations encode high-level concepts. There is a substantial body of work around the idea that concepts correspond to subspaces of a representation space (e.g., Mikolov et al., 2013a;b; Pennington et al., 2014; Goldberg & Levy, 2014; Arora et al., 2015; Gittens et al., 2017; Allen & Hospedales, 2019). Usually, this work focuses on a particular representation learning approach, and is either primarily empirical or offers a theoretical analysis closely tied to the particular domain of application. For example, in the context of word embeddings, theoretical explanations of the observed structure rely on the special structure of words and language (e.g., Arora et al., 2015; Allen & Hospedales, 2019). By contrast, the mathematical development in this paper is quite general—we only require that the data have two views separated by an underlying semantically meaningful space. We find that the concepts-as-subspaces structure is a general emergent phenomenon following from the structure of probability theory. It is not tied to any particular architecture or learning algorithm.

Our development also relates to a line of work that assumes the training data is generated by a particular latent variable model, and then shows that the learned representations (partially) recover the latent variables (e.g., Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Von Kügelgen et al., 2021; Eastwood et al., 2022; Higgins et al., 2018; Zimmermann et al., 2021). Often, a goal of this literature is to find representations that are "disentangled" in the sense that each dimension of the latent space corresponds to a single latent factor. In contrast, we do not assume a priori that there's a finite set of latent factors that generate the data. And, we view representations as defining probability distributions over latent concepts, rather than recovering the latent concepts themselves. As we have seen, this non-determinism is necessary in general.

Controlling Diffusion Models To demonstrate the concept-as-subspace structure, we developed a method for identifying the subspace corresponding to a given concept and showed how to manipulate concepts in the score representation of a diffusion model. We emphasize that our contribution here is not the manipulation procedure itself, but rather the mathematical framework that makes this procedure possible. In particular, the requirement to manipulate entire score functions is somewhat burdensome computationally. However, the ability to precisely manipulate individual concepts is clearly a useful tool, and it is an intriguing direction for future work to develop more efficient procedures for doing so. We conclude by surveying connections to existing work on controlling diffusion models.

One idea has been to take the bottleneck layer of UNet as a representation space and control the model by manipulating this space (Kwon et al., 2022; Haas et al., 2023; Park et al., 2023). This work does not consider text controlled models. It would be intriguing to understand the connection to the score-representation view, as moving from manipulation of the score to manipulation of the bottleneck layer would be a large computational saving.

Concept algebra can be seen as providing a unifying mathematical view on several methods that manipulate the score function (e.g., Du et al., 2021; Liu et al., 2021; Nair et al., 2022; Anonymous, 2023). Du et al. (2020); Liu et al. (2022) manipulate concepts via adding and subtracting scores. Negative prompting is a widely-used engineering trick that 'subtracts off' a prompt expressing unwanted concepts. Couairon et al. (2022) use score differences to identify objects' locations in images; this inspired our approach in Appendices B and E.1. In each case, we have seen that this kind of manipulation may be viewed as editing the subspace corresponding to some concept.

References

Allen, C. and Hospedales, T. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pp. 223–231. PMLR, 2019.

Anonymous. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OboQ71j1Bn. under review.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to pmi-based word embeddings, 2015.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Du, Y., Li, S., and Mordatch, I. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020.

Du, Y., Li, S., Sharma, Y., Tenenbaum, J., and Mordatch, I. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34: 15608–15620, 2021.

Eastwood, C., Nicolicioiu, A. L., von Kügelgen, J., Kekić, A., Träuble, F., Dittadi, A., and Schölkopf, B. Dci-es: An extended disentanglement framework with connections to identifiability. *arXiv preprint arXiv:2210.00364*, 2022.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Gittens, A., Achlioptas, D., and Mahoney, M. W. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76, 2017.

Goldberg, Y. and Levy, O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.

Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019.

Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., and Michaeli, T. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Hyvarinen, A. and Morioka, H. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Hyvarinen, A., Sasaki, H., and 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.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Liu, N., Li, S., Du, Y., Tenenbaum, J., and Torralba, A. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models, 2022.

Luo, C. Understanding diffusion models: A unified perspective, 2022.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013a.

Mikolov, T., Yih, W.-t., and Zweig, G. 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*, pp. 746–751, 2013b.

Mita, G., Filippone, M., and Michiardi, P. An identifiable double vae for disentangled representations. In *International Conference on Machine Learning*, pp. 7769–7779. PMLR, 2021.

Nair, N. G., Bandara, W. G. C., and Patel, V. M. Unite and conquer: Cross dataset multimodal synthesis using diffusion models, 2022.

Park, Y.-H., Kwon, M., Jo, J., and Uh, Y. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghase mipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35, 2022.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Zhou, D. and Wei, X.-X. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. *Advances in Neural Information Processing Systems*, 33:7234–7247, 2020.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A. Identifying Concept Subspace and its validity

A.1. Identifying the Concept Subspace

We estimate the concept space with Section 4.2. The justification for this procedure is based on the following proposition.

Proposition A.1. *Let Q_W be any fixed distribution over the W concept and Q_Z^0 be any reference distribution over Z. Then, assuming causal separability for $\mathcal{Z}, \mathcal{W}$,*

$$\mathcal{R}_Z(y) = \mathit{span}(\{s[Q_Z Q_W](y) - s[Q_Z^0 Q_W](y) : Q_Z \text{ a distribution}\}). \tag{8}$$

The importance of this expression is that it does not require the unknown s_Z. We can obtain elements of $\mathcal{R}_Z(y)$ with carefully chosen prompts such that $x_0, \dots x_k$ that elicit distributions of the form $Q_{x_j} = Q_Z^j Q_W$.

A.2. Validity

The procedure described above relies on finding spanning prompts $x_0, \dots, x_k$ for the target concept subspace. These prompts must satisfy $Q_{x_j} = Q_Z^j Q_W$ for some common Q_W, and we must have sufficient prompts to span the subspace. The first condition is a question of prompt design, and is often not too hard in practice. However, it is natural to wonder when it's possible to actually recover $\mathcal{R}_Z$ using only a practical number of prompts. We give some results showing that the dimension of $\mathcal{R}_Z(y)$ is often small, and thus can be spanned with a small number of prompts. Note that these results rely on the special structure of the score representation, and may not hold for other representations.

First, the case where $\mathcal{Z}$ is categorical with few categories:

Proposition A.2. *Assuming causal separability holds for $\mathcal{Z}, \mathcal{W}$. If $\mathcal{Z}$ is categorical with L possible values ($L \geq 2$), then $\mathit{dim}(\mathcal{R}_Z(y)) \leq L-1$.*

This result covers concepts such as `sex`, which can be spanned with only two prompts.

The next result extends this to certain categorical concepts with large cardinality, such as `style`. The idea is that if a concept is composed of finer grained categorical concepts, each with small cardinality, then the representation space of the concept is also low-dimensional. For example, `style` may be composed of lower-level concepts such as `color`, `stroke`, `textures`, etc.

Proposition A.3. *Suppose $\mathcal{Z}$ is composed of categorical concepts $\{\mathcal{Z}_k\}_{k=1}^K$ each with the number of categories L_k, in the sense that $\mathcal{Z} = \mathcal{Z}_1 \times \dots \mathcal{Z}_k$. Assume Y satisfies causal separability with respect to $\mathcal{Z}, \mathcal{W}$, with $Y_\mathcal{Z}$ the corresponding Y-measurable variable for $\mathcal{Z}$. Further assume that there exists $Y_\mathcal{Z}$-measurable variables $Y_{\mathcal{Z}_k}$ such that $p(y_\mathcal{Z} \mid z) = \Pi_{k=1}^K p(y_{\mathcal{Z}_k} \mid z_k)$. Then*

$$\mathit{dim}(\mathcal{R}_Z(y)) \leq \sum_{k=1}^K (L_k - 1) \tag{9}$$

Following this result, we might take the spanning prompts for style to be x_0 = "a mathematician in Art Deco style", x_1 = "a mathematician in Impressionist style", etc. Each of these prompts elicit a fixed distribution Q_W over the content, but varies the distribution Q_Z over style. If style is composed of finer-grained attributes, a relatively small set of such of prompts will suffice.

B. Examples

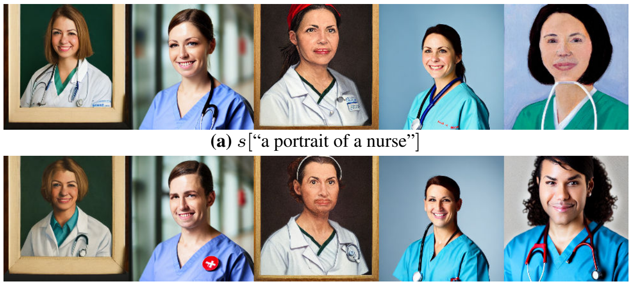
We have formalized what it means for concepts to correspond to subspaces of the representation space, and derived a procedure for identifying and editing the subspaces corresponding to particular concepts in the score representation. We now work through some examples testing if this subspace structure does indeed exist, and whether it enables manipulation of concepts.

Concepts We consider three concepts for our main experiments. First, the binary concepts `medium` = {`cartoon`, `photorealistic`}, concept `sex` = {`male`, `female`}. Following Proposition A.2, we need two prompts to elicit each of these spaces. Our choices are x_0^{sex} = "a man" and x_1^{sex} = "a woman" and x_0^{medium} = "cartoon" and x_1^{medium} = "photorealistic". Each of these prompts elicits a different distribution on the target concept, but a similar distribution on other concepts.

Next, we consider (artistic) `style` as a more complex concept. In this case, it is not practical to enumerate all possible styles. However, following Proposition A.3, we can hope to elicit the subspace with only a finite number of prompts.

There is an additional challenge here: in general, a prompt eliciting a particular style will also elicit other concepts. For example, adding the text "renaissance style" to a caption will tend to make any human subjects in the image be dressed in renaissance period clothing; see Figure 3c. That is, prompts x_0 = "renaissance style", x_1 = "postmodern style", x_3 = "impressionist style", ... don't meet the identification requirements of Proposition A.1. The reason is that the corresponding concept distributions are $Q_{x_j} = Q_Z^j Q_W^j$ with Q_W^j different for each j (e.g., the distribution over clothing styles is different for each style of art).

To overcome this, we choose prompts that are all forced to share the same Q_W distribution. Specifically, if we want to modify the prompt x we take our style prompts to be $x_0 = x$+"minimalist style", $x_1 = x$+"Japanese Ukiyo-e style",



(a) s[“a portrait of a nurse”]

(b) $(\mathbb{I}-\text{proj}_{\mathtt{sex}})s$[“a portrait of a nurse”]$+\text{proj}_{\mathtt{sex}}(\frac{1}{2}s$[“a female nurse”]$+\frac{1}{2}s$[“a male nurse”])

Figure 2. Elements of the representation subspace may not correspond to any prompt.

$x_3 = x+$“Impressionist style”, . . . In practice, we use ChatGPT to generate a list of 28 styles and use these.

Editing Concept Subspaces Given prompts defining each concept subspace, we construct the projection operation onto the subspace, and use this to produced edited representations. We use Stable Diffusion to sample from the distribution defined by the edited representations. We compare samples generated according to the original and edited representation, matching the random seed used for generation so that off-target concepts should match between samples. See Figures 1, 2 and 3 for results. In each case, we see that modifying the representation on the concept subspace leads to isolated modifications of the samples on the target concept.

Editing Concepts When Prompting Is Hard One interesting test is to see if we can edit a concept algebraically in situations where it is hard to modify that concept in isolation by prompting. Figure 3 shows examples of such cases. Here, modifying the style attribute with an English phrase (“photorealistic” or “in renaissance style”) elicits off-target behavior (e.g., changing the position of the frog, changing the background and clothing of the shoppers). However, the algebraic approach succeeds. This suggests that the internal representation can be readily manipulated, even when English prompting does not readily succeed.

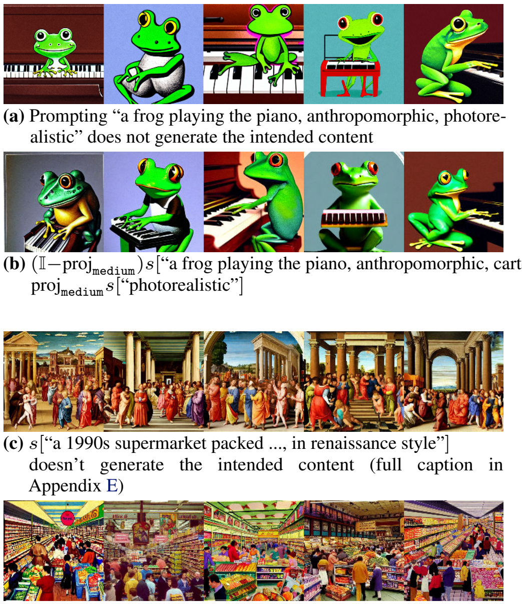
Promptless Concepts Another natural question is whether we can elicit concept values that cannot be achieved by direct prompting. In Figure 1b, we sample from

$$s_{\text{edit}} \leftarrow (\mathbb{I}-\text{proj}_{\mathtt{sex}})s[\text{x}]+\text{proj}_{\mathtt{sex}}s[\text{“person”}], \tag{10}$$

and observe that we eliminate the bias towards `male` elicited by the term “mathematician”. That is, we can generally elicit non-degenerate distributions on concepts. It is not possible to specify such behavior by direct prompting (because English doesn’t describe distributional aspects.) In Figure 2, we sample from

$$s_{\text{edit}} \leftarrow (\mathbb{I}-\text{proj}_{\mathtt{sex}})s[\text{x}] + \text{proj}_{\mathtt{sex}}\left(\frac{1}{2}s[\text{“a male nurse”}]+\frac{1}{2}s[\text{“a female nurse”}]\right). \tag{11}$$

The representation vector $\frac{1}{2}s$[“a male nurse”]$+\frac{1}{2}s$[“a female nurse”] does not correspond to any English prompt. We observe that modifications on the subspace still affect just the `sex` concept though—the samples are androgynous figures!



(a) Prompting “a frog playing the piano, anthropomorphic, photorealistic” does not generate the intended content

(b) $(\mathbb{I}-\text{proj}_{\mathtt{medium}})s$[“a frog playing the piano, anthropomorphic, cartoon”]$+\text{proj}_{\mathtt{medium}}s$[“photorealistic”]

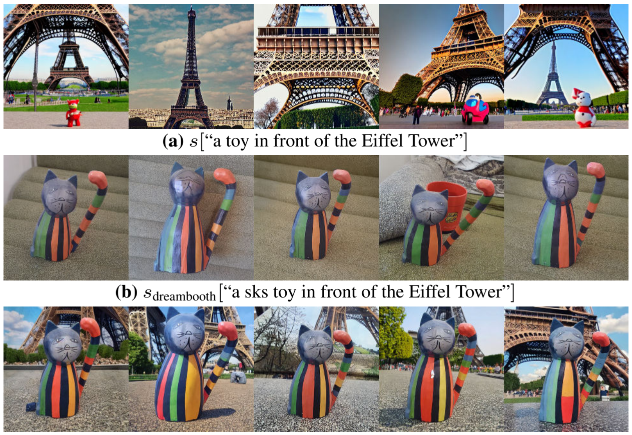
(c) s[“a 1990s supermarket packed ..., in renaissance style”] doesn’t generate the intended content (full caption in Appendix E)

(d) $(\mathbb{I}-\text{proj}_{\mathtt{style}})s$[“[content], in photorealistic style”]$+(\text{proj}_{\mathtt{style}})(s$[“[content], in renaissance-style painting”], with content = “a 1990s supermarket packed ...” (full caption in Appendix E)

Figure 3. Concept algebra can edit concepts in a disentangled fashion, even when direct prompting elicits off-target concepts

Mask as a Concept Finally, we consider a more abstract kind of concept motivated by the following problem. Suppose we have several photographs of a particular toy, and we want to generate an image of this toy in front of the Eiffel Tower. In principle, we can do this by fine-tuning the model (e.g., with dreambooth) to associate a new token (e.g., “sks toy”) with the toy. Then, we can generate the image by conditioning on the prompt “a sks toy in front of the Eiffel Tower”. In practice, however, this can be difficult because the fine-tuning ends up conflating the toy with the background in the demonstration images. E.g., the prompt “a sks toy in front of the Eiffel Tower” tends to generate images featuring carpet; see Figure 4b.

Intuitively, we might hope to fix this problem by finding a concept subspace that excludes background information. Given such a “subject subspace”, we could mask the subject



(a) s[“a toy in front of the Eiffel Tower”]

(b) $s_{\text{dreambooth}}$[“a sks toy in front of the Eiffel Tower”]

(c) $(\mathbb{I}-\text{proj}_{\texttt{subject}})s$[“a toy in front of the Eiffel Tower”] $+\text{proj}_{\texttt{subject}}s_{\text{dreambooth}}$[“a sks toy in front of the Eiffel Tower”]

Figure 4. We can manipulate abstract concepts such as ‘subject’ of the image

out of the image, generate the background, and then edit the subject back in. In Appendix E.1 we explain how to construct such a subspace using the prompts $x_0 =$ “a toy” and $x_1 =$ “a soccer ball”. Figure 4 shows the sampled output.

C. Concept Algebra Algorithms in Diffusion Model

Text-to-image Diffusion Models use score representations in their generation. More specifically, suppose the target is to sample $Y = Y_0 \sim P^*$, with the corresponding score function denoted as s_0. The key ingredients for generation are the score function for Y_t (denoted as s_t), which is Y noised at different levels, (e.g. $Y_t = (1-\alpha_t)Y+\alpha_t\epsilon_t$ for standard independent Gaussian noise ϵ), for $t = 0, ..., T$. See (Luo, 2022) for more details. To apply our results, we require causal separability with respect to $\mathcal{Z}, \mathcal{W}$ holds for all $Y_t, t = 0, ..., T$. Then our theoretical results follow through.

Algorithm 1 is an implementation of Concept Manipulation through Projection based on DDPM (Ho et al., 2020) (we can also implement different variants). [3] It requires FindSubspaceMethod, for which we can use FindSubspaceBasis(Algorithm 2) and FindSubspaceMask(Algorithm 3) based on the properties of $\mathcal{Z}$ as discussed in the main text. More specifically,

FindSubspaceBasis We calculate the projection matrix (denoted as Π_Z) for the Z-subspace, from a span of K prompts (after subtracting off the baseline) (Algorithm 2). In practical computations, we evaluate the $m\times K$ matrix $\triangle E$:

$$\triangle E := [\epsilon_\theta(y_t, t \mid x_1) - \epsilon_\theta(y_t, t \mid x_0), ..., \\ \epsilon_\theta(y_t, t \mid x_K) - \epsilon_\theta(y_t, t \mid x_0)]$$

Then, the top K_{thres} left singular vectors are selected as Q. Here, K_{thres} denotes the least number of factors required to surpass a certain proportion of variance explained, denoted as thres. Consequently, we have $\Pi_z \leftarrow QQ^T$.

FindSubspaceMask In this context, Π_Z signifies a mask. This mask can be calculated from the score difference $\triangle\epsilon$ (refer to Algorithm 3). As a practical measure, we may implement noise reduction techniques to fine-tune $\triangle\epsilon$. One approach is the application of a Gaussian blur to smooth out neighboring pixels.

Algorithm 1 Concept Manipulation through Projection

```
Require Diffusion model $\epsilon_\theta(y_t, t|x)$, guidance scale $w$, covariance matrix $\sigma_t^2 I$,
    empty prompt “”, prompts: $x_{\text{orig}}, x_{\text{new}}$,
    prompts to build the $Z$ subspace: $\{x_i\}$,
    the function for finding the $Z$ subspace: FindSubspaceMethod(·,·)
Initialize sample $y_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \ldots, 1$ do
    $\epsilon_{\text{empty}} \leftarrow \epsilon_\theta(y_t, t \mid “”)$ # unconditional score
    $\epsilon_{\text{orig}}, \epsilon_{\text{new}} \leftarrow \epsilon_\theta(y_t, t \mid x_{\text{orig}}), \epsilon_\theta(y_t, t \mid x_{\text{new}})$ # conditional scores
    $\Pi_Z \leftarrow$ FindSubspaceMethod$(y_t, \{x_i\})$ # find the projection matrix
    $\epsilon_{\text{cond}} \leftarrow (\mathbb{I}-\Pi_Z)\epsilon_{\text{orig}}+\Pi_Z\epsilon_{\text{new}}$ # concept projection
    $\epsilon \leftarrow \epsilon_0+w(\epsilon_{\text{cond}}-\epsilon_0)$ # apply classifier-free guidance
    $y_{t-1} \sim \mathcal{N}\left(y_t-\epsilon, \sigma_t^2 I\right)$
end for
```

Algorithm 2 FindSubspaceBasis

```
Require: $y_t \in \mathbb{R}^m$, prompts $\{x_k\}_{k=0}^K$
$\hat{\mathcal{R}_Z}(y) \leftarrow$ span$(\{\epsilon_\theta(y_t, t \mid x_k)-\epsilon_\theta(y_t, t \mid x_0)\}_{k=1}^K)$
Determine $\Pi_Z$ as the projection matrix onto $\hat{\mathcal{R}_Z}(y)$
return $\Pi_Z$
```

Algorithm 3 FindSubspaceMask

```
Require: $y_t \in \mathbb{R}^m$, a pair of prompts $(x_1, x_2)$
$\triangle\epsilon \leftarrow \epsilon_\theta(y_t, t \mid x_1)-\epsilon_\theta(y_t, t \mid x_2)$
for $i = 1$ to $m$ do
    $m_i \leftarrow \triangle\epsilon_i \neq 0?1:0$
end for
$\Pi_Z \leftarrow$ diag$(m_1, m_2, \ldots, m_m)$
return $\Pi_Z$
```

[3] Note there here we use residual $\epsilon_\theta(y_t, t \mid x)$ instead of the score $s_\theta(y_t, t \mid x)$ for generation, they are equivalent up to a time-varying constant.

D. Proofs

Proposition 3.1. *If Y is causally separable with respect to $\mathcal{W}$ and $\mathcal{Z}$, then the centered score representation is arithmetically composable with respect to $\mathcal{W}$ and $\mathcal{Z}$.*

Proof. By assumption in Definition 3.2, we have

$$\begin{aligned} p(y \mid z, w) &= p(y_{\mathcal{Z}}, y_{\mathcal{W}}, \xi(y) \mid z, w)\left|\det\left(\frac{\partial g}{\partial y}\right)\right| \\ &= p(y_{\mathcal{Z}} \mid z)p(y_{\mathcal{W}} \mid w)p(\xi(y))\left|\det\left(\frac{\partial g}{\partial y}\right)\right| \end{aligned}$$

Therefore,

$$\begin{aligned} p[Q](y) =& p[Q_Z \times Q_W](y) \\ =& p_Z[Q_Z](y)p[Q_W](y)p(\xi(y))\left|\det(\frac{\partial g}{\partial y})\right|, \end{aligned}$$

where $p_Z[Q_Z](y) = \int p(y_{\mathcal{Z}} \mid z)Q_Z(z)\mathrm{d}z$ and $p_W[Q_W](y) = \int p(y_{\mathcal{W}} \mid z)Q_W(w)\mathrm{d}w$.

Then, taking the log-derivative with respect to y, we get its score function as follows:

$$s[Q_Z \times Q_W](y) = s_Z[Q_Z](y) + s_W[Q_W](y) + s_0(y) \quad (12)$$

where $s_Z(y)$ and $s_W(y)$ are $p_Z[Q_Z](y)$'s and $p_W[Q_W](y)$'s score functions, and $s_0(y) := \nabla_y \log\left(p(\xi(y))\left|\det(\frac{\partial g}{\partial y})\right|\right)$. So the centered-score is

$$\begin{aligned} \bar{s}[Q_Z \times Q_W](y) =& (s_Z[Q_Z](y) - s_Z[Q_Z^0](y)) \\ &+ (s_W[Q_W](y) - s_W[Q_W^0](y)) \end{aligned}$$

where Q_Z^0 and Q_W^0 are the marginal distributions of Z and W of the baseline Q_0. Then, we can use the fact that

$$\begin{aligned} \mathcal{R}_Z &= \text{span}(\{\bar{s}_Z[Q_Z] - \bar{s}_Z[Q_Z^0] : Q_Z \text{ a distribution}\}) \\ &= \text{span}(\{s_Z[Q_Z] - s_Z[Q_Z^0] : Q_Z \text{ a distribution}\}) \\ \mathcal{R}_W &= \text{span}(\{\bar{s}_W[Q_W] - \bar{s}_W[Q_W^0] : Q_W \text{ a distribution}\}) \\ &= \text{span}(\{s_W[Q_W] - s_W[Q_W^0] : Q_W \text{ a distribution}\}) \end{aligned}$$

Consequently, the claim follows. □

Proposition A.1. *Let Q_W be any fixed distribution over the W concept and Q_Z^0 be any reference distribution over Z. Then, assuming causal separability for $\mathcal{Z}, \mathcal{W}$,*

$$\begin{aligned} \mathcal{R}_Z(y) =& \textit{span}(\{s[Q_Z Q_W](y) - s[Q_Z^0 Q_W](y) : \\ & Q_Z \textit{ a distribution}\}). \end{aligned} \quad (8)$$

Proof. By causal separability we can easily get the $\mathcal{R}_Z(y)$ in Proposition A.1 is the same as:

$$\mathcal{R}_Z(y) = \text{span}(\{s_Z[Q_Z](y) - s_Z[Q_Z^0](y)\} : Q_Z \text{ a distribution})$$

The only thing left to show is that $\mathcal{R}_Z(y)$ remains the same for whatever choice of baseline Q_Z^0. But this is immediate: $\text{span}(\{s_Z[Q_Z](y) - s_Z[Q_Z^0](y) : Q_Z \text{ a concept distribution}\}) = \text{span}(\{s_Z[Q_Z](y) - s_Z[Q_Z^1](y) : Q_Z \text{ a concept distribution}\})$ for any two baselines Q_Z^0 and Q_Z^1. □

Proposition A.2. *Assuming causal separability holds for $\mathcal{Z}, \mathcal{W}$. If $\mathcal{Z}$ is categorical with L possible values ($L \geq 2$), then $dim(\mathcal{R}_Z(y)) \leq L-1$.*

Proof. We denote the possible values that $\mathcal{Z}$ can take as $\{z_0, z_1, \ldots, z_{L-1}\}$. Let $\delta_{z_i} := \delta_{z_i}(z)$ represent the delta function in the $\mathcal{Z}$-subspace, which is infinite at z_i and zero at all other points. For any distribution Q_Z over Z and any $y \in \mathbb{R}^m$, we can express $s_Z[Q_Z](y)$ as a linear combination of $s_z[\delta_{z_i}]$ in the following form:

$$s_Z[Q_Z](y) = \sum_{l=0}^{L-1} \pi_l(y) s_Z[\delta_{z_l}](y)$$

Here, $\sum_{l=0}^{L-1} \pi_l(y) = 1$. Consider a baseline concept distribution Q_Z^0 and its corresponding Z-related score $s_Z[Q_Z^0](y) = \sum_{l=0}^{L-1} c_l(y) s_Z[\delta_{z_l}](y)$. We can then express the difference $s_Z[Q_Z](y) - s_Z[Q_Z^0](y)$ as:

$$s_Z[Q_Z](y) - s_Z[Q_Z^0](y) = \sum_{l=1}^{L-1} \omega_l(y)(s_z[\delta_{z_l}](y) - s_z[\delta_{z_0}](y)),$$

where $\omega_l(y) = \pi_l(y) - c_l(y)$ for $l = 1, \ldots, L-1$. Consequently, we can observe that $\mathcal{R}_Z(y) \subset \text{span}(\{s_z[\delta_{z_l}](y) - s_z[\delta_{z_0}](y)\}_{l=1}^{L-1})$, which implies that $\dim(\mathcal{R}_Z(y)) \leq L-1$. □

Proposition A.3. *Suppose $\mathcal{Z}$ is composed of categorical concepts $\{\mathcal{Z}_k\}_{k=1}^K$ each with the number of categories L_k, in the sense that $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \mathcal{Z}_k$. Assume Y satisfies causal separability with respect to $\mathcal{Z}, \mathcal{W}$, with $Y_{\mathcal{Z}}$ the corresponding Y-measurable variable for $\mathcal{Z}$. Further assume that there exists $Y_{\mathcal{Z}}$-measurable variables $Y_{\mathcal{Z}_k}$ such that $p(y_{\mathcal{Z}} \mid z) = \Pi_{k=1}^K p(y_{\mathcal{Z}_k} \mid z_k)$. Then*

$$dim(\mathcal{R}_Z(y)) \leq \sum_{k=1}^K (L_k - 1) \quad (9)$$

Proof. By assuming that $p(y_{\mathcal{Z}} \mid z) = \prod_{k=1}^K p(y_{\mathcal{Z}_k} \mid z_k)$, we can easily derive the following result for any concept distribution Q_Z over Z:

$$s_Z[Q_Z](y) = \sum_{k=1}^K s_{Z_k}[Q_{Z_k}](y),$$

where Q_{Z_k} represents the concept distribution of $\mathcal{Z}_k$ for each k, and $s_{Z_k}[Q_{Z_k}](y) = \nabla_y \log \left(\int p(y_{\mathcal{Z}_k} \mid z_k) Q_{Z_k}(z_k) \mathrm{d}z_k\right)$. Recall that

$$\mathcal{R}_Z(y) = \mathrm{span}(\{s_Z[Q_Z](y) - s_Z[Q_Z^0](y)\} : Q_Z \text{ is a concept distribution}),$$

where Q_Z^0 is a baseline. Importantly, it should be noted that $\mathcal{R}_Z(y)$ is unique regardless of the choice of Q_Z^0 as per Proposition A.1.

Let $Q_{Z_k}^0$ denote the $\mathcal{Z}_k$-related part of Q_Z^0 for $k = 1, \ldots, K$. We define $\mathcal{R}_{Z_k}(y) := \mathrm{span}(\{s_{Z_k}[Q_{Z_k}](y) - s_{Z_k}[Q_{Z_k}^0](y)\})$. Then, we can state that:

$$\mathcal{R}_Z(y) \subset \sum_{k=1}^{K} \mathcal{R}_{Z_k}(y).$$

Based on Proposition A.2, it follows that $\dim(\mathcal{R}_{Z_k}(y)) \leq L_k - 1$ for each k. Hence, we can conclude that:

$$\dim(\mathcal{R}_Z(y)) \leq \sum_{k=1}^{K} (L_k - 1).$$

□

E. Experiment Details and More Figures

E.1. Concept projection for Dreambooth (Figure 4)

First, we fine-tune the diffusion model using Dreambooth, applying a learning rate of $5e^{-6}$ and setting the number of steps to 800. While there are configurations that could yield a less overfitted model, we intentionally opt for these parameters to generate an overfitted model. Our aim is to verify if it's possible to disentangle the overfitted model by using concept manipulation via projection.

To generate images depicting `a sks toy in front of the Eiffel Tower`, we utilize our Dreambooth fine-tuned diffusion model together with the original pretrained Stable Diffusion model. Only for the new prompt, $x_{\text{new}} =$ a sks toy”, we use the score function from the Dreambooth fine-tuned model. All other prompts are plugged into the score functions from the original pretrained Stable Diffusion model. To create the desired images, we construct a projector using a pair of prompts: $(x_1, x_2) =$ (“a toy”, “a soccer ball”). The mask, computed using Algorithm 3 with the threshold$= 0.1$, helps identify specific areas corresponding to the location of the subject. Then, we use the Dreambooth score function $s_{\text{dreambooth}}$(a sks toy”) to guide the generation process within the masked region (areas with value 1), while using s(a toy in front of the Eiffel Tower”) to guide the generation outside the mask (areas with value 0).

To ensure image fidelity, we exclusively employ the score function $s_{\text{dreambooth}}$(“a sks toy”) for guiding the denoising process for the last 6% of the denoising steps.

It is important to note that due to severe overfitting issues with the fine-tuned model, there is no significant difference between using either the prompt “a sks toy” or “a sks toy in front of the Eiffel Tower” for the fine-tuned model. Also, due to the same reason, we apply the original pretrained diffusion model for all score functions except for the sks toy related one.

E.2. The mathematician example (Figure 1)

Our starting point is an original prompt $x_{\text{orig}} =$ “a portrait of a mathematician”. Our objective is to modify the `sex` and `style` using concept projection:

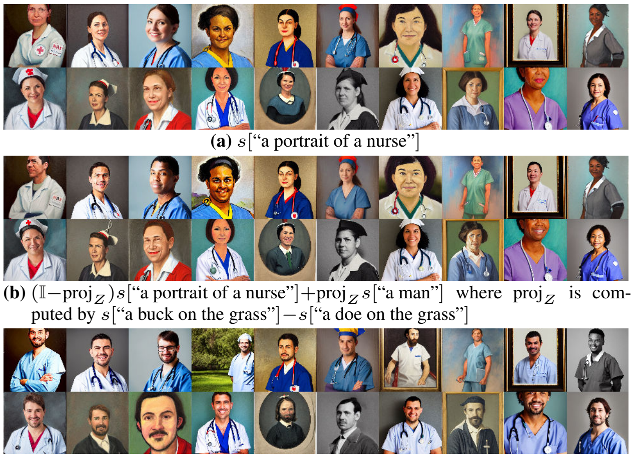
To adjust `sex`, we formulate a corresponding direction using a pair of prompts $(x_1, x_2) =$ (“a man”, “a woman”). Subsequently, we set $x_{\text{new}} =$ “a person”.

To alter the `style`, we set $x_{\text{new}} =$ “a portrait of a mathematician, in Fauvism style”. We define the concept subspace using prompts of the form “a portrait of a mathematician in [x_{style}] style”, where x_{style} takes value from a list of styles. During sampling, the original prompt is utilized in the first 20% of timesteps to better retain the content.

The list of styles is generated by ChatGPT. They are: Art Deco, Minimalist, Baroque, Abstract Expressionist, Cubist, Fauvism, Impressionist, Steampunk, Neoclassical, Japanese Ukiyo-e, Surrealism, Memphis Design, Scandinavian, Bauhaus, Pop Art, Art Nouveau, Street Art, American West, Victorian Gothic, Futurism, Photorealistic, Mannerist, Flemish, Byzantine, Medieval, Romanesque, Trompe-l’œil, and Dutch Golden Age.

E.3. The nurse example (Figure 2)

We initiate the process with an original prompt, $x_{\text{orig}} =$ “a portrait of a nurse”. Our goal is to perform concept projection to manipulate the `sex` attribute. Similar to the mathematician example, we define the `sex` direction using a pair of prompts: $(x_1, x_2) =$ (“a man”, “a woman”). However, instead of setting the distribution of `sex` as one of the delta functions or a fair one corresponding to the neutral prompt “a person”, we wish to see what the concept’s arithmetic average will define. Specifically, we take the `sex` direction of the average of `a female nurse` and `a male nurse`, calculated as $\frac{1}{2}s$(a female nurse”)$+\frac{1}{2}s$(a male nurse”). It turns out the arithmetic mean realize the interpolation between two extremal sex points in the `sex` subspace, and the score function after concept projection returns images of androgynous nurses.



(a) s[“a portrait of a nurse”]

(b) $(\mathbb{I}-\text{proj}_Z)s$[“a portrait of a nurse”]$+\text{proj}_Z s$[“a man”] where proj_Z is computed by s[“a buck on the grass”]$-s$[“a doe on the grass”]

(c) $(\mathbb{I}-\text{proj}_Z)s$[“a portrait of a nurse”]$+\text{proj}_Z s$[“a man”] where proj_Z is computed by s[“a man”]$-s$[“a woman”]

Figure 5. Necessity of Assumptions: the validity of concept algebra depends on causal separability

E.4. Failure of direct prompting (Figure 3)

Two instances are presented where direct prompting does not yield the desired outcome. We circumvent this limitation by merging x_{orig} and x_{new} via projection:

In the case of the frog, using the prompt “a frog playing the piano, anthropomorphic, photorealistic” fails to yield the desired content-style combination. To rectify this, we initially set $x_{\text{orig}} =$ “a frog playing the piano, anthropomorphic, cartoon” to obtain the desired content. Then, we adjust the `medium` to `photorealistic` by setting $x_{\text{new}} =$ “photorealistic”. We determine $\text{proj}_{\text{medium}}$ using the pair $(x_1, x_2) =$ (“cartoon”, “photorealistic”).

For the mall scenario, the content described as

> $x_{\text{content}} =$ “a 1990s supermarket packed to the brim with people, showcasing a lively, shoulder-to-shoulder shopping experience.”

If we add “a renaissance-style of” to x_{content} in the prompt, the resulting images often contain mythological figures common in renaissance-style paintings. However, this can be avoided by combining x_{orig} and x_{new} via projection. Here, x_{orig} is defined as a photo of the content x_{content} and x_{new} is “a renaissance-style of” appended to the x_{content}. The `style` subspace is identified using basis prompts of the format “x_{content} in x_{style} style”, with x_{style} using the same list of files styles generated by ChatGPT, as in Figure 1.

F. Additional experiments

We show that our methods would fail when causal separability assumption (Definition 3.2) become invalid. In this section, we show one concrete example of failures.

Figure 5 shows that we are unable to transfer the gender of the nurse when we calculate the score function of `a male nurse` by $(\mathbb{I}-\text{proj}_Z)s$[“a portrait of a nurse”]$+\text{proj}_Z s$[“a man”] where proj_Z is computed by s[“a buck on the grass”]$-s$[“a doe on the grass”]. The target concept Z and W are `sex` $\in$ {`male, female`} and `species`$\in$ {`human, deer`}. It’s obvious that the `sex` and `species` have an interaction effect on the image Y — different species induce different sexual characteristics.