MULTIMODAL DATASETS WITH CONTROLLABLE MUTUAL INFORMATION

Raheem Karim Hashmani
University of Wisconsin–Madison
hashmani@wisc.edu

Garrett W. Merz
University of Wisconsin–Madison
garrett.merz@wisc.edu

Helen Qu
Flatiron Institute
hqu@flatironinstitute.org

Mariel Pettee
University of Wisconsin–Madison
mpettee@wisc.edu

Kyle Cranmer
University of Wisconsin–Madison

ABSTRACT

We introduce a framework for generating highly multimodal datasets with explicitly calculable mutual information (MI) between modalities. This enables the construction of benchmark datasets that provide a novel testbed for systematic studies of mutual information estimators and multimodal self-supervised learning (SSL) techniques. Our framework constructs realistic datasets with known MI using a flow-based generative model and a structured causal framework for generating correlated latent variables. We benchmark a suite of MI estimators on datasets with varying ground truth MI values and verify that regression performance improves as the MI increases between input modalities and the target value. Finally, we describe how our framework can be applied to contexts including multi-detector astrophysics and SSL studies in the highly multimodal regime.[1]

1 INTRODUCTION

Self-supervised learning (SSL) has become a core component of many state-of-the-art large-scale machine learning models (Balestriero et al., 2023). Such models are also increasingly *multimodal*, i.e. designed to learn from varied input sources such as text, images, and audio (Radford et al., 2021; Zong et al., 2024). A prevailing intuition is that multimodal SSL is effective because different modalities provide complementary "views" of the same underlying concepts, enabling the learning process to exploit their shared information. The precise relationship between the mutual information (MI) between modalities and SSL performance, however, is not fully understood.

Contrastive SSL methods built using the InfoNCE loss function (Oord et al., 2018; Chen et al., 2020; He et al., 2020; Caron et al., 2020) have a clear information-theoretic interpretation: for example, the learned similarity scores estimate the pointwise mutual information (PMI) between paired samples. By contrast, no analogous theoretical connection has been established for either highly multimodal settings (i.e. $N > 2$ modalities) or for non-contrastive SSL methods such as multimodal masked modeling (Mizrahi et al., 2023; He et al., 2022; Wang et al., 2023; Huang et al., 2025), despite the fact that both of these directions are quickly gaining prominence in the field (Li et al., 2024; Hondru et al., 2025). A theoretically-grounded understanding of the fundamental relationship between inter-modality mutual information and SSL representations (and their corresponding performance on downstream tasks) will be increasingly critical as models continue to scale to larger numbers of input modalities. In particular, principled frameworks will be needed to evaluate how the distribution of shared information across modalities influences the quality of the learned embeddings.

[1]Our code is publicly available at https://github.com/RKHashmani/MmMi-Datasets.

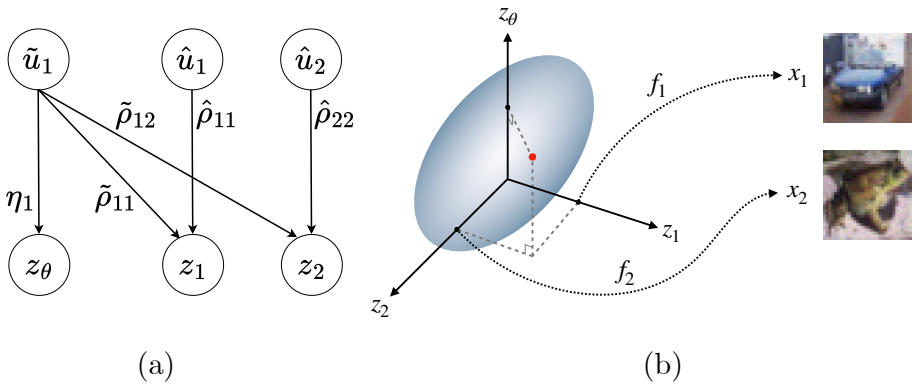


Figure 1: **Schematic of our dataset generation framework.** a) An example DAG showing the linear mixing of proto-latents $\mathbf{u}$ via coefficients η, ρ into interpretable correlated Gaussian latent variables $\mathbf{z}$. b) Overview of sampling from a multidimensional Gaussian to draw latent inputs z_1, z_2, and z_θ that are fed into invertible maps f_1 and f_2 to a realistic feature space.

Complicating matters further, MI is notoriously difficult to estimate from samples, particularly in high-dimensional, real-world datasets (McAllester & Stratos, 2020; Czyż et al., 2023a). A wide range of MI estimators have been proposed using techniques such as kernel estimation, k-nearest neighbor, and neural estimators (Pizer et al., 1987; Kozachenko & Leonenko, 1987; Moon et al., 1995; Kraskov et al., 2004; Belghazi et al., 2018; Song & Ermon, 2020b; Butakov et al., 2024; Belghazi et al., 2021). However, these estimators are typically only validated on synthetic datasets of simple distributions for which the MI is analytically tractable (Darbellay & Vajda, 1999; Suzuki, 2016; Czyż et al., 2023a;b; Butakov et al., 2023).

Datasets with controllable MI that emulate the challenges of real-world data are needed to better understand the advantages, disadvantages, and tradeoffs of different SSL learning objectives. Such datasets can enable systematic, reproducible studies of how multimodal SSL representations depend on information overlap and shared features, offering both theoretical insights and practical guidance for model design. Finally, they provide a reliable testbed for evaluating the performance of various mutual information estimation strategies designed for use on real-world datasets.

In this work, we introduce a framework to generate realistic multimodal data with controllable mutual information. Figure 1 shows an overview of our data generation framework[2]:

(a) First, we use directed acyclic graphs (DAGs) to generate easily interpretable correlated Gaussian latent variables $\mathbf{z}$ with known mutual information.

(b) We then feed the outputs $\mathbf{z}$ of these DAGs into invertible bijective transformations to construct multimodal datasets where the amount and distribution of shared information can be explicitly controlled across multiple modalities.

2 Background

Mutual Information (MI). Mutual information $I(X;Y)$ is a fundamental quantity from information theory that measures the statistical dependence between two random variables X and Y. It is formally defined as the Kullback-Leibler (KL) divergence of the joint distribution $p(X,Y)$ and the product of the marginal distributions $p(X)p(Y)$:

$$I(X;Y) \equiv D_{KL}\left(p(X,Y) \mid\mid p(X)p(Y)\right)$$

Alternatively, it can also be expressed in terms of the Pointwise Mutual Information (PMI):

$$I(X;Y) \equiv \mathbb{E}_{x,y\sim p(X,Y)}\left[\text{PMI}(x;y)\right]$$

[2]Datasets generated based on Figure 1 can be found at (Hashmani et al., 2026).

The MI quantifies the extent to which one variable reduces uncertainty in the other. For instance, when X and Y are fully independent, their joint distribution $p(X, Y)$ reduces to the product of the marginal distributions $p(X)p(Y)$, therefore the MI is zero.

Pointwise Mutual Information (PMI). When evaluated on specific values $x \sim p(X)$ and $y \sim p(Y)$, the pointwise MI (PMI) captures the probability of x and y occurring together compared with that same probability if they were fully independent. The PMI is formally expressed as:

$$\text{PMI}(x; y) \equiv \log \left(\frac{p(x, y)}{p(x)p(y)} \right).$$

Multimodal Self-Supervised Learning (SSL). Often compared to the primary human senses such as sight, hearing, or touch, modalities in machine learning refer to distinct forms of sensing the world and the corresponding representations of the observed data. Modalities can have radically different formats (e.g. RGB images and timeseries data), or they can exhibit similar formats but describe distinct information sources (e.g. RGB images and segmentation maps). In this paper, our operational definition for a modality is a random variable X_m and a corresponding sample space $\mathcal{X}_m$. In particular, if X_1 and X_2 correspond to two different distributions, then we consider them to be two separate modalities regardless of their data format. Multimodal SSL often involves learning a joint representation of many data modalities, which we view as distinct from multi-*view* SSL, which generally consists of learning a joint representation of multiple views derived from the same data modality, e.g. different crops of a single image. Multimodal SSL uses the relationships between modalities to learn joint representations without explicit labels.

Flow-based generative modeling. Flow-based generative models are designed to facilitate the direct transformation between probability distributions using invertible mappings applied to a simple base distribution such as a Gaussian. Each transformation is designed to be bijective with a tractable Jacobian determinant, enabling exact computation of both likelihoods and samples. Because they provide both efficient sampling and exact density evaluation, flow-based models are increasingly used not only in generative modeling but also in scientific applications requiring tractable likelihoods and explicit control over distributions. *Flow-matching* (Albergo & Vanden-Eijnden, 2022; Lipman et al., 2024) is a recent approach within the family of flow-based generative models that enables efficient training of Continuous Normalizing Flows (CNFs) (Chen et al., 2018a) by directly regressing the velocity field that transports a base distribution to the data distribution instead of optimizing the exact maximum-likelihood objective.

3 Creating Datasets with Controlled Mutual Information

Our goal is to enable rigorous, scalable experiments using multimodal datasets where the MI between modalities is precisely specified and easy to interpret. To accomplish this, we design an expressive three-step framework $\mathbf{u} \rightarrow \mathbf{z} \rightarrow \mathbf{x}$. This begins with uncorrelated, normally distributed 'proto-latent' variables $\mathbf{u}$, which are related by linear structural equations to form an easy-to-interpret causal model for latent variables $\mathbf{z}$, for which mutual information is easy to compute. Finally, we use blocks of components of $\mathbf{z}$ as the input to a set of invertible transformations $\{f_i\}_{i=1}^n$ (one for each of n modalities) to produce synthetic observations $\mathbf{x}_i = f_i(\mathbf{z}_i)$ that preserve the mutual information between the corresponding latent variables. In this work, we implement f_i as flow-matching models that have been pretrained to produce realistic images.

In addition to the $\mathbf{x}_i$ for each modality, we also generate a (scalar) target variable θ computed from the latent variable z_θ. We partition the vector of proto-latents into sets of components $\mathbf{u} = (\tilde{\mathbf{u}}, \hat{\mathbf{u}})^T$. The goal here is to isolate a source of randomness $\tilde{\mathbf{u}}$ that can be interpreted as a common cause that induces correlation between the observed $\mathbf{x}_i$ and some target quantity of interest θ that one may wish to estimate from the $\mathbf{x}_i$. For simplicity, we take θ to be a scalar and let $\theta = z_\theta$ since complicated non-linear relationships between $\mathbf{x}_i$ and θ are already captured by the flows f_i.

3.1 Generalized Linear Causal Construction: Proto-latent to Latent Connections

We wish to create a large latent vector $\mathbf{z}$ that is distributed according to a multivariate Gaussian with known covariance for which the mutual information is easy to compute. We achieve this with linear combinations of i.i.d. normally distributed proto-latents $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{z} = (z_\theta, \mathbf{z}_1, \ldots, \mathbf{z}_{N_z})^\top = \mathbf{A}\,(\tilde{\mathbf{u}}, \hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_{N_u})^\top \tag{1}$$

where:

- $\tilde{\mathbf{u}} \in \mathbb{R}^{N_\theta}$: proto-latents serving as a common cause inducing correlation between θ and $\mathbf{x}_i$,
- $\hat{\mathbf{u}}_j \in \mathbb{R}^d$: proto-latents for the observed modalities, $j = 1, \ldots, N_u$,
- $z_\theta \in \mathbb{R}$: scalar target quantity of interest (e.g., a physical quantity to be estimated from $\mathbf{x}_i$),
- $\mathbf{z}_i \in \mathbb{R}^d$: latent variables associated to each observed modality, $i = 1, \ldots, N_z$,
- $\mathbf{A}$: a matrix specifying the structural equations in the causal model relating $\mathbf{u}$ and $\mathbf{z}$.

The matrix $\mathbf{A}$ encodes all structured dependencies between latent variables and outputs. One could use an arbitrary matrix $\mathbf{A}$, but that would lack interpretability. Instead, we structure $\mathbf{A}$ to follow from an expressive, easy-to-interpret causal story.

For example, the causal model shown in Fig.1(a) corresponds to the structural equations

$$z_\theta = \eta \tilde{u}_1 \;; \qquad \mathbf{z}_1 = \tilde{\rho}_{11} \tilde{u}_1 \, \mathbf{1}_d + \hat{\rho}_{11} \hat{\mathbf{u}}_1 \;; \qquad \mathbf{z}_2 = \tilde{\rho}_{12} \tilde{u}_1 \, \mathbf{1}_d + \hat{\rho}_{22} \hat{\mathbf{u}}_2 \;, \tag{2}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. The hyperparameters of this model are $\eta, \tilde{\rho}_{ki}, \hat{\rho}_{ji} \in \mathbb{R}$. Note we treat $\tilde{\mathbf{u}}$ and $\hat{\mathbf{u}}$ asymmetrically: $\tilde{\mathbf{u}}$ is a common cause that feeds into both z_θ and the latents associated to the individual modalities, while $\hat{\mathbf{u}}$ does not feed into z_θ. In this simple example, $\hat{\mathbf{u}}$ is also the only source of correlation between z_1 and z_2 – and the mutual information between z_1 and z_2 is perfectly predictive of z_θ.

In Section 4 we will consider other causal stories, their corresponding linear structural equations, and the consequences of these relationships on the induced mutual information between θ and $\mathbf{x}_i$.

3.2 Analytic Mutual Information of the Latent Variables

We provide a derivation of the mutual information calculation between latent variables $\mathbf{z}$ constructed as described in Section 3.1. The covariance matrix of the latents is simply given by:

$$\Sigma = \mathrm{Cov}(\mathbf{Z}, \mathbf{Z}) = \mathbf{A}\mathbf{A}^\top \tag{3}$$

We can represent the covariance matrix in block form corresponding to z_θ, $\mathbf{z}_1$, and $\mathbf{z}_2$ as Σ (and for any two blocks in Σ, we can define the reduced block matrix Γ) as:

$$\Sigma = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta 1} & \Sigma_{\theta 2} \\ \Sigma_{1\theta} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{2\theta} & \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad \Gamma_{ij} = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \tag{4}$$

For multivariate Gaussian distributions, the mutual information is a simple function of the determinants of these block covariance matrices. For example,

$$I(\theta; Z_1) = \frac{1}{2} \ln \left(\frac{|\Sigma_{\theta\theta}||\Sigma_{11}|}{|\Gamma_{\theta 1}|} \right) \qquad I(Z_1; Z_2) = \frac{1}{2} \ln \left(\frac{|\Sigma_{11}||\Sigma_{22}|}{|\Gamma_{12}|} \right) \;, \tag{5}$$

where $|\cdot|$ denotes the determinant of the corresponding block covariance matrix.

While the covariance matrix and mutual information quantities in the preceding equations can be calculated numerically, we are also able to derive closed-form, analytical equations for various mutual information quantities (see Appendix B). One benefit of the closed form solutions is they reveal

scaling in terms of the hyperparameters of the structural equations, the number of modalities, and the dimensonality of each modality. For the causal model considered in Fig. 1(a) and Eq. 2, we find:

$$I(\theta; Z_1) = -\frac{1}{2}\log\left(1 - \frac{d\,\tilde{\rho}_{11}^2}{\hat{\rho}_{11}^2 + d\,\tilde{\rho}_{11}^2}\right) \tag{6}$$

$$I(Z_1; Z_2) = -\frac{1}{2}\log\left(1 - \frac{d^2\,\tilde{\rho}_{11}^2\tilde{\rho}_{12}^2}{[\hat{\rho}_{11}^2 + d\,\tilde{\rho}_{11}^2][\hat{\rho}_{22}^2 + d\,\tilde{\rho}_{12}^2]}\right)\ . \tag{7}$$

These equations that we have derived have been verified against the numerical calculations.

3.3 Flow-based generative modeling preserves mutual information

The final step of our three-step process is to create realistic synthetic data in multiple modalities. Recall that the latent vector is organized by blocks of components as $\mathbf{z} = (z_\theta, \mathbf{z}_1, \ldots, \mathbf{z}_{N_l})^T$. We transform the individual blocks of latent variables independently, yielding $\mathbf{x}_i = f_i(\mathbf{z}_i)$, where the $f_i(\cdot)$ are generative models pretrained on real-world datasets.

We leverage a key result that states that if the f_i are continuous bijective maps, then the mutual information is preserved:

$$I(X_i; X_j) = I(Z_i; Z_j)\ . \tag{8}$$

This result can be seen as following from the data-processing inequality and is also the result of a direct computation of the mutual information after a change of variables, where the Jacobian factors that arise cancel exactly (see e.g., Cover & Thomas, 2006; Czyż et al., 2023a;b). While many generative models satisfy this condition, e.g., discrete-time normalizing flows (Rezende & Mohamed, 2015; Papamakarios et al., 2021), we use continuous-time normalizing flows based on flow matching (Lipman et al., 2024; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022) in this work. We pretrain on CIFAR-10 (Krizhevsky, 2009) using image class as a proxy for modality (i.e., f_0 is trained on images of cars, f_1 is trained on images of frogs), but we emphasize that our framework is agnostic to f_i parameterization and modality definition.

3.4 Templates

The mutual information $I(X_i, X_j)$ does not specify how this information is distributed across the components of X_i and X_j. Similarly, the pointwise mutual information in two images does not uniquely determine the spatial location of their correlated pixels. Nevertheless, the way the information is distributed matters in practice because architectural choices are sensitive to those details. The impact of these architectural choices on the performance of competing approaches to SSL or mutual information estimation then become conflated other algorithmic choices (e.g. data augmentation and training objectives) that are more clearly tied to (pointwise) mutual information.

Ideally, we would like to perform ablation studies designed to disentangle these effects. This requires being able to independently vary the mutual information and the way that information is distributed across the components of the random variables. In order to achieve this, we introduce the notion of *templates* into our $\mathbf{u} \to \mathbf{z}$ mapping.

We define a template $\mathbf{T}_{ik} \in \mathbb{R}^d$ as a linear map relating the common cause $\tilde{u}_k$ and the a latent $\mathbf{z}_i$:

$$\mathbf{z}_i = \sum_{k=1}^{N_\theta} \tilde{u}_k \mathbf{T}_{ik} + \sum_{j=1}^{N_u} \hat{\mathbf{u}}_j \tag{9}$$

For example, $\mathbf{T}_{ik} = \frac{1}{d}\mathbf{1}_d$ implements a homogeneous distribution of information about $\tilde{u}_k$ across the latent $\mathbf{z}_i$, while $\mathbf{T}_{ik} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ implements a scenario where all the information about $\tilde{u}_k$ is concentrated in a single component of $\mathbf{z}_i$.

This design is motivated by real-world scenarios in which the information about multiple common causes is distributed nonuniformly across several modalities (e.g., multiple supernovae being imaged by multiple types of telescopes). With templates, future studies can better understand the impact of architectural design choices based on the information distribution in various modalities.

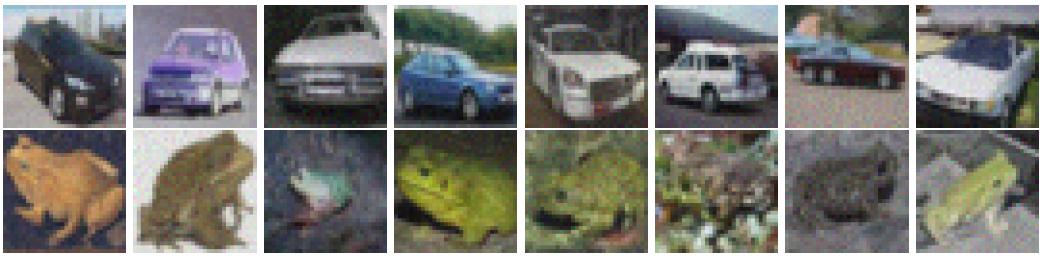


Figure 2: Eight examples of correlated pairs of images $(\mathbf{x}_1, \mathbf{x}_2)$ generated from our procedure representing two realistic modalities. In this case, both modalities are images, but corresponding to flows conditioned on different class labels ("automobile", "frog") from CIFAR-10 (Krizhevsky, 2009). The dimensionality of the data in both cases is $d = 32 \times 32 \times 3 = 3072$.

4 Examples of datasets resulting from our model

As mentioned in Sec. 3.1, the matrix $\mathbf{A}$ allows for an arbitrary linear structural equation between the proto-latents $\mathbf{u}$ and the latent variables $\mathbf{z}$. While this flexibility may come at the cost of interpretability, we find that in fact many realistic causal stories are well captured by structural equations with only a few hyperparameters.

We show two specific examples in this section. All examples are implemented using a (conditional) flow matching model pretrained on CIFAR-10 data, where image class label is used as a proxy for different modalities. Figure 2 shows eight examples of correlated pairs $(\mathbf{x}_1, \mathbf{x}_2)$ generated from our procedure. **While there is no clear visual connection between these pairs of images, our framework allows us to state unequivocally that these high-dimensional, complex image pairs have a specific quantity of mutual information – a feat that was previously unattainable.**

4.1 Example 1: Empirically Demonstrating Example Use Cases

Benchmarking mutual information estimators. We demonstrate the use of our framework to generate a benchmark dataset to explore the performance of a number of popular mutual information estimators. We use the causal model shown in Figure 1 to generate a set of ten datasets with mutual information $I(X_1; X_2)$ ranging from 0.0284 to 1.39, with each dataset consisting of 10,000 paired CIFAR-like images. We use an existing benchmark suite (Lee & Rhee, 2024) to estimate the empirical mutual information and compare it to the ground-truth mutual information in Figure 3. We report the correlation and RMSE for each estimator in Table 1. We observe that the regressed mutual information from all the estimators follows the linearly increasing ground truth mutual information.

MI Estimator	Correlation	RMSE
DV (Donsker & Varadhan, 1983)	0.995	0.1094
JS (Nowozin et al., 2016)	0.993	0.3049
InfoNCE (Oord et al., 2018)	0.991	0.1981
MINE (Belghazi et al., 2018)	0.993	0.0983
NWJ (Nguyen et al., 2010)	0.996	0.0851
SMILE-1 (Song & Ermon, 2020a)	0.998	0.0935
SMILE-5 (Song & Ermon, 2020a)	0.993	0.0974
SMILE-inf (Song & Ermon, 2020a)	0.993	0.0969

Table 1: Correlation and RMSE from linear regression of estimated vs. ground-truth mutual information, computed for a number of mutual information estimators.

Regressing the target variable θ from X. We evaluate how well a model is able to regress the target variable θ from data X_1 as a function of the mutual information $I(\theta, X_1)$. Intuitively, models with fixed capacity should predict θ more accurately from data that contain more information about θ, i.e. larger $I(\theta; X_1)$. We re-use the same ten datasets, for which $I(\theta; X_1)$ ranges from 0.134 to

1.73. For each dataset, we regress θ from the images using a shallow convolutional network, choosing the best model after training for 500 epochs. We show that the best achievable RMSE decreases with increasing mutual information between X_1 and θ (Figure 4) and show example distributions of prediction error $\theta - \hat{\theta}$ for two representative MI values in Appendix A.

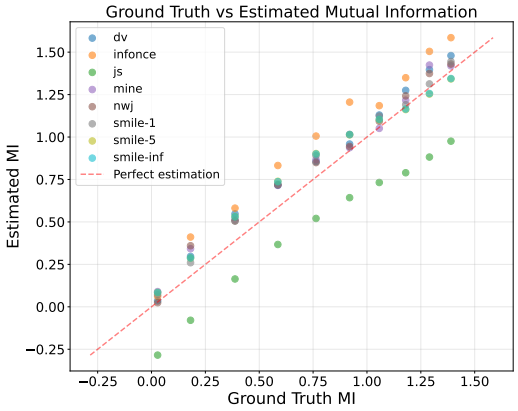


Figure 3: MI estimators vs. ground-truth MI.

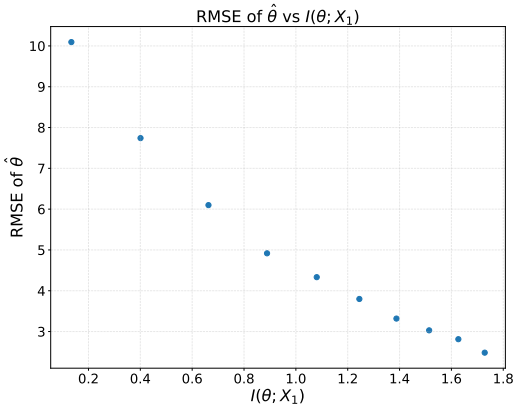
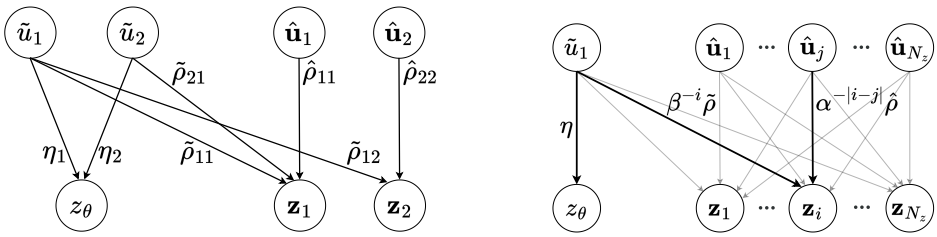


Figure 4: Regression of a target scalar θ monotonically improves with increasing $I(X_1, \theta)$.

4.2 Example 2: Estimating a black hole's mass from two telescopes

Consider the hypothetical scenario where one wishes to estimate the mass of Sagittarius A*, the supermassive black hole in the center of our Milky Way Galaxy. To do this we might employ two instruments producing two data modalities. Let $\mathbf{x}_1$ represent data from the Event Horizon Telescope, a ground-based array consisting of a global network of radio telescopes. Let $\mathbf{x}_2$ represent data from the Hubble space telescope in orbit around the Earth. Let $\tilde{u}_1$ represent the unknown mass of the black hole and let $\tilde{u}_2$ represent some atmospheric variability that impacts how radio waves propagate.



(a) Causal structure for black hole example.

(b) Causal structure for multimodal example.

Figure 5: Example causal structures with corresponding structural equations that induce specific MI.

The mass of the black hole $\tilde{u}_1$ influences the data from both telescopes; however, the atmospheric effects $\tilde{u}_2$ only impact the data from the Event Horizon Telescope. This narrative is captured by the causal model illustrated in Fig. 5(a). This causal model corresponds to the structural equations

$$z_\theta = \eta_1 \tilde{u}_1 + \eta_2 \tilde{u}_2 \; ; \quad \mathbf{z}_1 = \tilde{\rho}_{11}\tilde{u}_1 \mathbf{T}_{11} + \tilde{\rho}_{12}\tilde{u}_2 \mathbf{T}_{12} + \hat{\rho}_{11}\hat{\mathbf{u}}_1 \; ; \quad \mathbf{z}_2 = \tilde{\rho}_{21}\tilde{u}_1 \mathbf{T}_{21} + \hat{\rho}_{22}\hat{\mathbf{u}}_2 \; . \quad (10)$$

The closed-form, analytical equations for various MI quantities corresponding to a similar causal model can be found in Appendix B. The $\mathbf{T}_{ki}$ are *templates* that have the same shape as the $\mathbf{z}_i$ and can encode some type of inhomogeneous (spatial) structure in the latents. For example, the templates $\mathbf{T}_{11}$ and $\mathbf{T}_{12}$ associated with $\tilde{u}_1$ (black hole mass) are designed to concentrate at the center of the galaxy and dissipates away from the center. Similarly, the template $\mathbf{T}_{21}$ associated with $\tilde{u}_2$ (atmospheric effect) are designed to be diffuse across the whole example.

Fig. 5(a) also has paths from $\tilde{u}_1$ and $\tilde{u}_2$ to z_θ scaled by the hyperparameters η_1 and η_2. This flexibility allows us to capture two different narratives in the same model by changing the values of η_i. In one

narrative, θ represents the mass of the black hole and corresponds to $\eta_1 = 1, \eta_2 = 0$. In the second narrative, θ represents the atmospheric effect and corresponds to $\eta_1 = 0, \eta_2 = 1$.

Table 2 shows the result of the mutual information when all of the ρ variables are set to 1 and the dimensionality of the data in each modality is $d = 3072$. Note that in the scenario where the quantity of interest θ corresponds to the atmospheric effect, that there is no mutual information between the data from the Event Horizon Telescope and the quantity of interest.

Table 2: Mutual information for two scenarios corresponding to the causal structure in Fig. 5(a).

θ Represents	η_1 (Black hole)	η_2 (Atmosphere)	$I(\theta; X_1)$	$I(\theta; X_2)$	$I(X_1; X_2)$
Black Hole Mass	1	0	2.77	0	2.63
Atmospheric Effect	0	1	2.77	3.33	2.63

4.3 Example 3: A scalable model for massively multimodal data

In this example, we shift our emphasis to the number of modalities. The ability to generate correlated tuples of synthetic data $(\mathbf{x}_1, \ldots, \mathbf{x}_{N_z})$ with known mutual information will be extremely valuable for studying the tradeoff among various competing approaches to multimodal SSL. We would like a flexible template that allows us to generate a large number of modalities while keeping a small, fixed number of hyperparameters to reason about. At the same time, we would like the model to be expressive enough to capture some interesting patterns.

The causal model illustrated in Fig. 5(b) corresponds to the structural equations

$$z_\theta = \eta \tilde{u}_1 \; ; \qquad \mathbf{z}_i = \beta^{-i} \tilde{\rho}\, \tilde{u}_1 \mathbf{1} + \sum_{j=1}^{N_u} \alpha^{-|i-j|} \hat{\rho}\, \hat{\mathbf{u}}_j \; . \tag{11}$$

Each set of proto-latents $\hat{\mathbf{u}}_i$ has a corresponding set of latents $\mathbf{z}_i$, which they feed into with a single coefficient $\hat{\rho}$. In addition, the jth proto-latents also contribute to the ith latents with some decay constant $\alpha^{-|i-j|}$, with $\alpha \geq 1$. As the hyper-parameter α grows, the correlation between the modalities decays quickly (as a function of $|i - j|$). As $\alpha \to 1$, the modalities become uniformly correlated.

Here we maintain a target quantity of interest for some downstream task (e.g. regression), but only include a single common cause $\tilde{u}_1$. This common cause also induces a correlation among the $\mathbf{z}_i$, but we break the permutation invariance by including a scaling factor β^{-i}. When β is large, only the first few modalities have significant mutual information with θ; however, when $\beta \to 1$, that mutual information with θ is uniform.

This simple model does not reflect a specific physical scenario, but it does allow for interesting benchmarks and experiments for multimodal SSL. We show in Figure 6 results from training a flow-matching model on 10 CIFAR class labels, allowing us to create these correlated tuples of high-dimensional, realistic images for up to $N_z = 10$. Specifically, we show the mutual information when all of the ρ variables are set to 1, the dimensionality of the data in each modality is $d = 3072$, and various α and β are selected. We note that as α increases, $I(X_1; X_i)$ decays at a faster rate and as β increases, $I(\theta; X_i)$ decays at a faster rate, as expected. Extending beyond 10 modalities is a straightforward exercise.

4.4 Example 4: A model for ablation studies for multimodal SSL

While the example in Sec. 4.3 allows one to study the performance of multi-modal SSL methods as a function of the mutual information between the modalities (and the pointwise mutual information between individual samples from those modalities), it does not provide a mechanism to probe the impact of architectural choices on the performance of various methods. Different architectural choices can be sensitive to the distribution of information across the feature components of a modality (e.g. how the information is distributed across pixels in an image). In this example we introduce structural equations that enable ablation studies that can independently isolate the role of (pointwise) mutual information from the distribution of information.

As discussed in Section 3.4, templates can control how the information is distributed among the components of each latent $\mathbf{z}_i$ while preserving the total mutual information between two modalities.

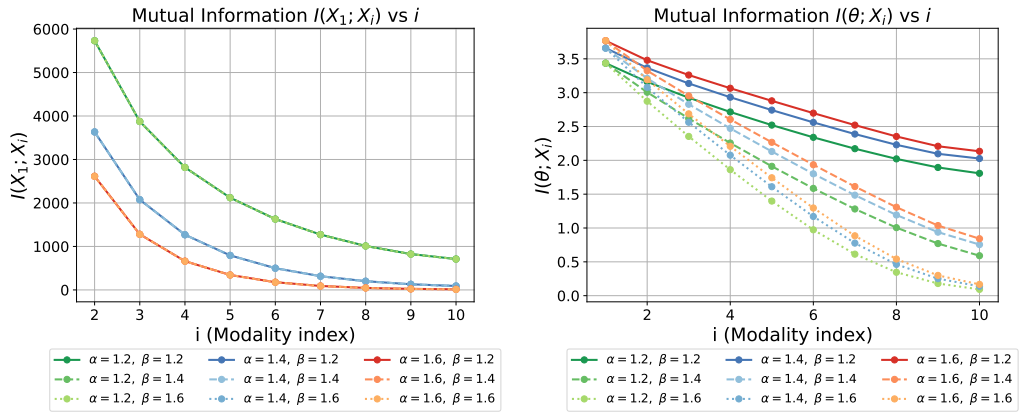


Figure 6: As expected, $I(X_1, X_i)$ (left) and $I(\theta, X_i)$ (right) decrease with the index i.

This provides a mechanism for disentangling the effects of algorithmic (e.g. specific SSL objectives) from architectural choices (e.g. inductive biases in the model construction) by independently varying how the information is distributed in the data. The following equations incorporate modality-specific templates associated to a set of shared proto-latents representing common causes $\tilde{u}_k$ as well as a path for shared information from the proto-latents $\hat{\mathbf{u}}_j$ that are independent of the target latent z_θ:

$$z_\theta = \sum_{k=1}^{N_\theta} \eta_k \tilde{u}_k \qquad \mathbf{z}_i = \sum_{k=1}^{N_\theta} \tilde{\rho}_{ik} \tilde{u}_k \mathbf{T}_{ik} + \sum_{j=1}^{N_u} \hat{\rho}_{ij} \hat{\mathbf{u}}_j, \tag{12}$$

The coefficients $\tilde{\rho}_{ik}$ and $\hat{\rho}_{ij}$ could either be independent hyperparameters or they could be parametrized as in Section 4.3, e.g. $\tilde{\rho}_{ik} = \beta^{-i}\tilde{\rho}$ and $\hat{\rho}_{ij} = \alpha^{-|i-j|}\hat{\rho}$.

5 Related Work

Self-supervised learning and mutual information. The relevance of mutual information (MI) to self-supervised learning (SSL), particularly contrastive learning, has been studied extensively. For example, the InfoNCE family of contrastive loss functions can be interpreted as bounds on the MI between representations (Oord et al., 2018; He et al., 2020; Caron et al., 2020). A range of work inspired by the InfoMax principle (Linsker, 1988) has argued that the MI between inputs and learned representations is an implicit target for multiview contrastive learning (Oord et al., 2018; Hjelm et al., 2019; Hénaff et al., 2020; Tsai et al., 2021). However, Tschannen et al. (2020) find that maximizing the MI alone is not sufficient for learning representations that are useful for downstream tasks, stressing that the relation between estimated MI and representation quality depends strongly on both architecture choice and the form of the MI estimator used. More recent work (Tian et al., 2020; Wang et al., 2022; Wang & Isola, 2022; Rodríguez-Gálvez et al., 2023) explores this question in further depth. Our ability to generate realistic, complex datasets with known MI may enable further progress in determining the role of information maximization in self-supervised learning.

Mutual information estimation and benchmarking. Estimating mutual information from samples is challenging (McAllester & Stratos, 2020), especially in compelling real-world datasets and applications (Holmes & Nemenman, 2019; Gao et al., 2015; 2017). A rich body of literature covers a variety of mutual information estimation methods, ranging from traditional approaches based on histogram density or k-nearest neighbors (Pizer et al., 1987; Kozachenko & Leonenko, 1987; Kraskov et al., 2004) to neural estimators based on variational approaches (Belghazi et al., 2018; Song & Ermon, 2020b) or generative modeling (Ao & Li, 2022; Butakov et al., 2024). However, benchmarking the efficacy of these estimators on realistic datasets is entirely nontrivial. Most existing approaches are validated on multivariate normal distributions where mutual information is easily controllable, while recent work has explored simple transformations of these distributions to emulate properties of real data (Czyż et al., 2023a;b; Butakov et al., 2023). Our work replaces these simple

transformations with a flexible bijective mapping learned through flow-based generative modeling (Chen et al., 2018b), enabling construction of highly realistic datasets with analytically tractable MI.

Mutual information-preserving transforms. Several generative models satisfy the mutual information preservation condition. Discrete-time normalizing flows with coupling layers (Rezende & Mohamed, 2015; Papamakarios et al., 2021) and Invertible Residual Networks (Behrmann et al., 2019) were among the first invertible (bijective) deep generative models. More recently, TARFLOW and STARFLOW (Zhai et al., 2024; Gu et al., 2025) have achieved very strong results on high resolution image generation. The ability of flow-based models to preserve mutual information has been employed in a variety of contexts, including mutual information estimation (Butakov et al., 2024) and developing alternative prescriptions for training flow models (Ardizzone et al., 2021). However, employing this property to develop realistic datasets with known mutual information remains a novel contribution of this work.

6 CONCLUSION

We present a new framework for generating realistic datasets with many modalities that are designed with known and controllable mutual information. Our dataset generation framework uses interpretable causal models with linear structural equations to construct correlated, normally-distributed latent variables with known mutual information. Blocks of components of these random variables are then fed into invertible (bijective) transformations that map the latent inputs into a realistic feature space while preserving the mutual information content. These realistic and nontrivial datasets enable numerous studies, including benchmarking studies of mutual information estimators. Critically, these datasets will be important for understanding the role of mutual information in various multimodal self-supervised learning strategies, particularly as the number of modalities grows.

REPRODUCIBILITY

To encourage reproducibility, we submit all code as supplementary material. Additionally, we describe the details of our example datasets, including hyperparameters chosen. Upon publication, we will release the code and sample datasets publicly.

ACKNOWLEDGMENTS

MP gratefully acknowledges the Center for Computational Astrophysics at the Flatiron Institute for hospitality while a portion of this work was carried out. GM and KC are supported in part by the U.S. Department of Energy (DOE) under Award No. DE-FOA-0002705, KA/OR55/22 (AIHEP). This research was also supported in part by grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP), at which a portion of this work was performed.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *ArXiv*, abs/2209.15571, 2022. URL https://api.semanticscholar.org/CorpusID:252668615.

Ziqiao Ao and Jinglai Li. Entropy estimation via normalizing flow. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9990–9998, 2022.

Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training normalizing flows with the information bottleneck for competitive generative classification, 2021. URL https://arxiv.org/abs/2001.06448.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Joern-Henrik Jacobsen. Invertible residual networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2021. URL https://arxiv.org/abs/1801.04062.

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, Alexey Frolov, and Kirill Andreev. Information bottleneck analysis of deep neural networks via lossy compression. *arXiv preprint arXiv:2305.08013*, 2023.

Ivan Butakov, Aleksandr Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, and Alexey Frolov. Mutual information estimation via normalizing flows. *Advances in Neural Information Processing Systems*, 37:3027–3057, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Changyou Chen, Chunyuan Li, Liqun Chen, Wenlin Wang, Yunchen Pu, and Lawrence Carin. Continuous-time flows for efficient inference and density estimation, 2018a. URL https://arxiv.org/abs/1709.01179.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Paweł Czyż, Frederic Grabowski, Julia Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. *Advances in neural information processing systems*, 36:16957–16990, 2023a.

Paweł Czyż, Frederic Grabowski, Julia E Vogt, Niko Beerenwinkel, and Alexander Marx. On the properties and estimation of pointwise mutual information profiles. *arXiv preprint arXiv:2310.10240*, 2023b.

Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 36:183–212, 1983.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 277–286. PMLR, 2015.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Advances in Neural Information Processing Systems*, pp. 5986–5997, 2017.

Jiatao Gu, Tianrong Chen, David Berthelot, Huangjie Zheng, Yuyang Wang, Ruixiang Zhang, Laurent Dinh, Miguel Angel Bautista, Josh Susskind, and Shuangfei Zhai. Starflow: Scaling latent normalizing flows for high-resolution image synthesis. *arXiv preprint arXiv:2506.06276*, 2025.

Raheem Hashmani, Garrett Merz, Helen Qu, Mariel Pettee, and Kyle Cranmer. mmmi-dag1-2modalities-cifar10 (revision 6b5d362), 2026. URL https://huggingface.co/datasets/RKHashmani/mmmi-dag1-2modalities-cifar10.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019. URL https://arxiv.org/abs/1808.06670.

Caroline M Holmes and Ilya Nemenman. Estimation of mutual information for real-valued data with error bars and controlled bias. *Physical Review E*, 100(2):022404, 2019.

Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey, 2025. URL https://arxiv.org/abs/2408.06687.

Tao Huang, Yanxiang Ma, Shan You, and Chang Xu. Learning mask invariant mutual information for masked image modeling. *arXiv preprint arXiv:2502.19718*, 2025.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2020. URL https://arxiv.org/abs/1905.09272.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23:9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009.

Kyungeun Lee and Wonjong Rhee. A benchmark suite for evaluating neural mutual information estimators on unstructured datasets, 2024. URL https://arxiv.org/abs/2410.10924.

Siyuan Li, Luyuan Zhang, Zedong Wang, Di Wu, Lirong Wu, Zicheng Liu, Jun Xia, Cheng Tan, Yang Liu, Baigui Sun, and Stan Z. Li. Masked modeling for self-supervised representation learning on vision and beyond, 2024. URL https://arxiv.org/abs/2401.00897.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, March 1988. ISSN 0018-9162. doi: 10.1109/2.36. URL https://doi.org/10.1109/2.36.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mcallester20a.html.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, 52:2318–2321, Sep 1995. doi: 10.1103/PhysRevE.52.2318. URL https://link.aps.org/doi/10.1103/PhysRevE.52.2318.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, November 2010. ISSN 1557-9654. doi: 10.1109/tit.2010.2068870. URL http://dx.doi.org/10.1109/TIT.2010.2068870.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 271–279, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Borja Rodríguez-Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning, 2023. URL https://arxiv.org/abs/2307.10907.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. 2020a.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020b.

Joe Suzuki. An estimator of mutual information and its application to independence testing. *Entropy*, 18(4):109, 2016.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning?, 2020. URL https://arxiv.org/abs/2005.10243.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective, 2021. URL https://arxiv.org/abs/2006.05576.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning, 2020. URL https://arxiv.org/abs/1907.13625.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning, 2022. URL https://arxiv.org/abs/2203.07004.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022. URL https://arxiv.org/abs/2005.10242.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey, 2024. URL https://arxiv.org/abs/2304.01008.

A Prediction Error for the θ Regression Task

In Figure 7, we show the distribution of prediction errors $\theta - \hat{\theta}$ for two illustrative points along our MI distribution. The dataset with lower MI (in blue) exhibits a wider spread in errors, indicating decreased regression performance, while the dataset with higher MI (in orange) shows a sharper peak, indicating better regression performance. Moreover, both distributions are centered at 0, indicating that they are not biased in their estimations of θ.

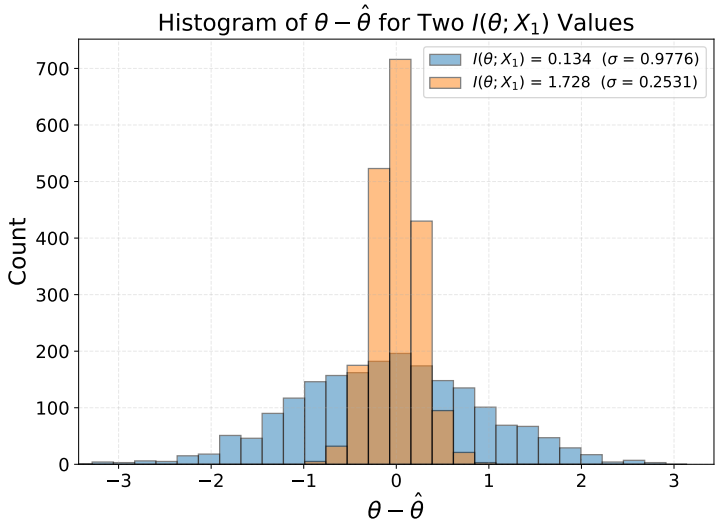


Figure 7: An example distribution of prediction error $\theta - \hat{\theta}$ for two representative MI values.

B Analytic formulae for covariance matrices and mutual information

For a causal story where:

$$
\begin{aligned}
z_\theta &= \sum_{k=1}^{N_\theta} \eta_k \cdot \tilde{u}_k \\
\mathbf{z}_i &= \sum_{k=1}^{N_\theta} \tilde{\rho}_{ik} \cdot \tilde{u}_i + \sum_{j=1}^{N_u} \hat{\rho}_{ij} \cdot \hat{\mathbf{u}}_i
\end{aligned}
\tag{13}
$$

We can represent the covariance matrix for a bimodal case as follows:

$$
\Sigma = \begin{bmatrix} a & r_1 \mathbf{1}^\top & r_2 \mathbf{1}^\top \\ r_1 \mathbf{1} & (b-o)\mathbf{I}_d + o\mathbf{J}_d & (f-p)\mathbf{I}_d + p\mathbf{J}_d \\ r_2 \mathbf{1} & (f-p)\mathbf{I}_d + p\mathbf{J}_d & (c-q)\mathbf{I}_d + q\mathbf{J}_d \end{bmatrix}
\tag{14}
$$

where $\mathbf{I}_d$ and $\mathbf{J}_d$ are the $d \times d$ identity and matrix of ones, respectively, and:

- $a = \sum_{k=1}^{N_\theta} \eta_k^2$
- $r_1 = \sum_{k=1}^{N_\theta} \eta_k \, \tilde{\rho}_{1k} \qquad r_2 = \sum_{k=1}^{N_\theta} \eta_k \, \tilde{\rho}_{2k}$

- $o = \sum_{k=1}^{N_\theta} (\tilde{\rho}_{1k})^2 \qquad q = \sum_{k=1}^{N_\theta} (\tilde{\rho}_{2k})^2$
- $p = \sum_{k=1}^{N_\theta} \tilde{\rho}_{1k}\tilde{\rho}_{2k}$
- $b = o + \sum_{j=1}^{N_u} (\hat{\rho}_{1j})^2 \qquad c = q + \sum_{j=1}^{N_u} (\hat{\rho}_{2j})^2$
- $f = p + \sum_{j=1}^{N_u} \hat{\rho}_{1j}\hat{\rho}_{2j}$

Because $\mathbf{I}_d$ and $\mathbf{J}_d$ commute, they can be simultaneously diagonalized. Thus, for a block like

$$\Sigma_{11} = (b-o)\mathbf{I}_d + o\mathbf{J}_d, \tag{15}$$

the eigenvalues are:

- $b + (d-1)o$ (multiplicity 1, for $\mathbf{1}_d$)
- $b - o$ (multiplicity $d-1$, for vectors orthogonal to $\mathbf{1}_d$).

Therefore,

$$|\Sigma_{11}| = (b-o)^{d-1}[b + (d-1)o] \tag{16}$$

$$|\Sigma_{22}| = (c-q)^{d-1}[c + (d-1)q\ . \tag{17}$$

By the matrix determinant lemma and Schur complement, the determinant of a block matrix Γ_{ij} can be written as

$$|\Gamma_{ij}| = |\Sigma_{jj}| \cdot \left|\Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}\right|\ , \tag{18}$$

and thus

$$|\Gamma_{\theta 1}| = (b-o)^{d-1}\left[a(b + (d-1)o) - dr_1^2\right] \tag{19}$$

$$|\Gamma_{12}| = \left[(b-o)(c-q) - (f-p)^2\right]^{d-1}\left[(b + (d-1)o)(c + (d-1)q) - (f + (d-1)p)^2\right]\ . \tag{20}$$

Some examples of closed-form equations using these terms are

$$I(\theta; Z_1) = -\frac{1}{2}\log\left(1 - \frac{dr_1^2}{a[b + (d-1)o]}\right) \tag{21}$$

$$I(\theta; Z_2) = -\frac{1}{2}\log\left(1 - \frac{dr_2^2}{a[c + (d-1)q]}\right) \tag{22}$$

$$\begin{aligned} I(Z_1; Z_2) = &\frac{d-1}{2}\log\left(\frac{(b-o)(c-q)}{(b-o)(c-q) - (f-p)^2}\right) \\ &+ \frac{1}{2}\log\left(\frac{[b + (d-1)o][c + (d-1)q]}{(b + (d-1)o)(c + (d-1)q) - (f + (d-1)p)^2}\right)\ . \end{aligned} \tag{23}$$

These equations have been verified against the numerical calculations.