IDENTIFIABILITY RESULTS FOR MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning is a cornerstone underlying recent progress in multi-view and multimodal learning, e.g., in representation learning with image/caption pairs. While its effectiveness is not yet fully understood, a line of recent work reveals that contrastive learning can invert the data generating process and recover ground truth latent factors shared between views. In this work, we present new identifiability results for multimodal contrastive learning, showing that it is possible to recover shared factors in a more general setup than the multi-view setting studied previously. Specifically, we distinguish between the multi-view setting with one generative mechanism (e.g., multiple cameras of the same type) and the multimodal setting that is characterized by *distinct* mechanisms (e.g., cameras and microphones). Our work generalizes previous identifiability results by redefining the generative process in terms of distinct mechanisms with modality-specific latent variables. We prove that contrastive learning can block-identify latent factors shared between modalities, even when there are nontrivial dependencies between factors. We empirically verify our identifiability results with numerical simulations and corroborate our findings on a complex multimodal dataset of image/text pairs. Zooming out, our work provides a theoretical basis for multimodal representation learning and explains in which settings multimodal contrastive learning can be effective in practice.

1 INTRODUCTION

Multimodal representation learning is an emerging field whose growth is fueled by recent developments in weakly-supervised learning algorithms and by the collection of suitable multimodal datasets. Multimodal data is characterized by the *co-occurence* of observations from two or more dependent data sources, such as paired images and captions (e.g., Salakhutdinov and Hinton, 2009; Shi et al., 2019; Radford et al., 2021), and more generally, multimodal observations are comprised of aligned measurements from different types of sensors (Baltrušaitis et al., 2019). Co-occurrence is a form of *weak supervision* (Shu et al., 2020; Locatello et al., 2020; Chen and Batmanghelich, 2020), in that paired observations can be viewed as proxies (i.e., weak labels) for a shared but unobserved ground truth factor. Among suitable representation learning methods for weakly supervised data, *contrastive learning* (Gutmann and Hyvärinen, 2010; Oord et al., 2018) stands out because it is designed to leverage co-occurring observations from different views. In practice, contrastive learning achieves promising results for multi-view and multimodal learning—a prominent example is the contribution of `CLIP` (Radford et al., 2021) to the groundbreaking advancements in text-to-image generation (Ramesh et al., 2021; 2022; Rombach et al., 2022; Saharia et al., 2022).

Despite its empirical success, it is not sufficiently well understood what explains the effectiveness of contrastive learning in practice. Recent works attribute its effectiveness to the recovery of shared latent factors from the underlying causal graph (Gresele et al., 2019; Zimmermann et al., 2021; von Kügelgen et al., 2021). From the perspective of multi-view independent component analysis, it was shown that contrastive learning can invert a nonlinear mixing function (i.e., a nonlinear generative process) that is applied to a latent variable with mutually independent components (Gresele et al., 2019; Zimmermann et al., 2021). More recently, von Kügelgen et al. (2021) show that contrastive learning can recover shared factors up to block-wise indeterminacies, even if there are nontrivial causal and statistical dependencies between latent components. Collectively, these results suggest that contrastive learning can identify parts of an unknown data generating process from pairs of observations alone—even from

high-dimensional multi-view observations with nontrivial dependencies. In our work, we investigate the identifiability of shared latent factors in the *multimodal* setting.

We consider a generative process with modality-specific mixing functions and modality-specific latent variables. Our design is motivated by the inherent heterogeneity of multimodal data, which follows naturally when observations are generated by different types of sensors (Baltrušaitis et al., 2019). For example, an agent can perceive its environment through distinct sensory modalities, such as cameras sensing light or microphones detecting sound waves. To model information that is shared between modalities, we take inspiration from the multi-view setting (von Kügelgen et al., 2021) and allow for nontrivial dependencies between latent variables. However, previous work only considers observations of the same data type and assumes that the same input leads to the same output across views. In this work, we introduce a model with *distinct generative mechanisms*, each of which can exhibit a significant degree of modality-specific variation. This distinction renders the multimodal setting more general compared to the multi-view setting considered by previous work.

In a nutshell, our work is concerned with *identifiability for multimodal representation learning* and focuses on *contrastive learning* as a particular algorithm for which we derive identifiability results. In Section 2, we cover relevant background on both topics, identifiability and contrastive learning. We then formalize the multimodal generative process as a latent variable model (Section 3) and prove that contrastive learning can block-identify latent factors shared between modalities (Section 4). We empirically verify the identifiability results with fully controlled numerical simulations (Section 5.1) and corroborate our findings on a complex multimodal dataset of image/text pairs (Section 5.2). Finally, we contextualize related literature (Section 6) and discuss potential limitations and opportunities for future work (Section 7).

2 Preliminaries

2.1 Identifiability

Identifiability lies at the heart of many problems in the fields of independent component analysis (ICA), causal discovery, and inverse problems, among others (Lehmann and Casella, 2006). From the perspective of ICA, we consider the relation $\mathbf{x} = \mathbf{f}(\mathbf{z})$, where an observation $\mathbf{x}$ is generated from a mixing function $\mathbf{f}$ that is applied to a latent variable $\mathbf{z}$. The goal of ICA is to invert the mixing function in order to recover the latent variable *from observations alone*. In many settings, full identifiability is impossible and certain ambiguities might be acceptable. For example, identifiability might hold for a subset of components (i.e., partial identifiability). Typical ambiguities include permutation and element-wise transformations (i.e., component-wise indeterminacy), or identifiability up to groups of latent variables (i.e., block-wise indeterminacy). In the general case, when $\mathbf{f}$ is a nonlinear function, a landmark negative result states that the recovery of the latent variable given i.i.d. observations is fundamentally impossible (Hyvärinen and Pajunen, 1999). However, a recent line of pioneering works provides identifiability results for the difficult nonlinear case under additional assumptions, such as auxiliary variables (Hyvärinen and Morioka, 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020) or multiple views (Gresele et al., 2019; Locatello et al., 2020; Zimmermann et al., 2021).

Most relevant to our investigation are previous works related to *multi-view nonlinear ICA* (Gresele et al., 2019; Lyu and Fu, 2020; Locatello et al., 2020; von Kügelgen et al., 2021). Generally, this line of work considers the following generative process:

$$\mathbf{z} \sim p_{\mathbf{z}}, \quad \mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}), \tag{1}$$

where a latent variable, or a subset of its components, is shared between *pairs* of observations $(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1,\mathbf{x}_2}$, where the two views $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by two nonlinear mixing functions, $\mathbf{f}_1$ and $\mathbf{f}_2$ respectively. Intuitively, a second view can resolve ambiguity introduced by the nonlinear mixing, if both views contain a shared signal but are otherwise sufficiently distinct (Gresele et al., 2019). Previous works differ in their assumptions on the mixing functions and dependence relations between latent components. The majority of previous work (Song et al., 2014; Gresele et al., 2019; Lyu and Fu, 2020; Locatello et al., 2020) considers mutually independent latent components, and

some of these works (Song et al., 2014; Gresele et al., 2019; Lyu and Fu, 2020) consider view-specific[1] mixing functions. Venturing beyond the strict assumption of independent components, von Kügelgen et al. (2021) consider additional causal and statistical dependencies between latent components and show that the subset of shared components can be identified up to a block-wise indeterminacy. Our work considers heterogeneous modalities with causal and statistical dependencies between latents. We prove that shared factors can be block-identified in a novel setting with modality-specific mixing functions and modality-specific latent variables.

2.2 Contrastive Learning

Contrastive learning (Gutmann and Hyvärinen, 2010; Oord et al., 2018) is a self-supervised representation learning method that leverages weak supervision in the form of paired observations. On a high level, the method learns to distinguish "positive" pairs of encodings sampled from the joint distribution, against "negative" pairs sampled from the product of marginals. The popular InfoNCE objective (Oord et al., 2018) is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{\{\mathbf{x}_1^i, \mathbf{x}_2^i\}_{i=1}^K \sim p_{\mathbf{x}_1, \mathbf{x}_2}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{g}_1(\mathbf{x}_1^i), \mathbf{g}_2(\mathbf{x}_2^i))/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\mathbf{g}_1(\mathbf{x}_1^i), \mathbf{g}_2(\mathbf{x}_2^j))/\tau\}} \right], \quad (2)$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are encoders for the first and second view, $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. It is common to use a single encoder $\mathbf{g}_1 = \mathbf{g}_2$ when $\mathbf{x}_2$ is an augmented version of $\mathbf{x}_1$ or when two augmentations are sampled from the same distribution to transform $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively (e.g., Chen et al., 2020). The set of hyperparameters consists of the temperature τ, a similarity metric $\text{sim}(\cdot, \cdot)$, and an integer K that controls the number of negative pairs ($K - 1$) used for contrasting. The objective has an information-theoretic interpretation as a variational lower bound on the mutual information $I(\mathbf{g}_1(\mathbf{x}_1); \mathbf{g}_2(\mathbf{x}_2))$ (Oord et al., 2018; Poole et al., 2019) and it can also be interpreted as the alignment of positive pairs (numerator) with additional entropy regularization (denominator), where the regularizer disincentivizes a degenerate solution in which both encoders map to a constant (Wang and Isola, 2020). Formally, when instantiating the $\mathcal{L}_{\text{InfoNCE}}$ objective with $\tau = 1$ and $\text{sim}(a, b) = -(a - b)^2$, it asymptotically behaves like

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) = \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1, \mathbf{x}_2}} \left[\|\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\|_2 \right] - H(\mathbf{g}(\mathbf{x})) \quad (3)$$

for a single encoder $\mathbf{g}$, when $K \to \infty$ (Wang and Isola, 2020; von Kügelgen et al., 2021).

In the setting with two heterogeneous modalities, it is natural to employ separate encoders $\mathbf{g}_1 \neq \mathbf{g}_2$, which can represent different architectures. Further, it is common to use a symmetrized version of the objective (e.g., see Zhang et al., 2020; Radford et al., 2021), which can be obtained by computing the mean of the loss in both directions:

$$\mathcal{L}_{\text{SymInfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) = {}^1\!/_2\, \mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2) + {}^1\!/_2\, \mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_2, \mathbf{g}_1). \quad (4)$$

Akin to Equation (3), we can approximate the symmetrized objective for $\tau = 1$ and $\text{sim}(a, b) = -(a - b)^2$, with a large number of negative samples ($K \to \infty$), as follows:

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1, \mathbf{x}_2}} \left[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2 \right] - {}^1\!/_2 \left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2)) \right). \quad (5)$$

Since the similarity measure is symmetric, the approximation of the alignment term is identical for both $\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1, \mathbf{g}_2)$ and $\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_2, \mathbf{g}_1)$. Each entropy term is approximated via the denominator of the respective loss term, which can be viewed as a nonparametric entropy estimator (Wang and Isola, 2020). For our experiments, we employ the finite-sample estimators $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{SymInfoNCE}}$, while for our theoretical analysis we use the estimand $\mathcal{L}_{\text{SymAlignMaxEnt}}$ to derive identifiability results.

3 Generative Process

In the following, we formulate the multimodal generative process as a latent variable model (Section 3.1) and then specify our technical assumptions on the relation between modalities (Section 3.2).

[1] Note that we define *modality-specific* functions similar to the way Gresele et al. (2019) and Lyu and Fu (2020) define *view-specific* functions. To clarify the distinction, we generally assume that observations from different modalities are generated by distinct mechanisms $\mathbf{f}_1 \neq \mathbf{f}_2$ with modality-specific latent variables, and we treat the multi-view setting as a special case, where $\mathbf{f}_1 = \mathbf{f}_2$ without view-specific latent variables.

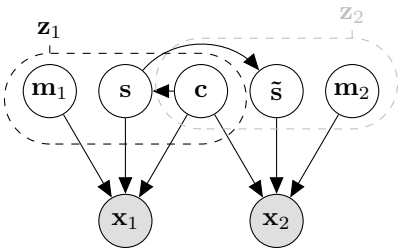


Figure 1: Illustration of the multimodal generative process. Latent variables are denoted by clear nodes and observations by shaded nodes. We partition the latent space into $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 = (\tilde{\mathbf{c}}, \tilde{\mathbf{s}}, \mathbf{m}_2)$, where $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 1) and hence we consider only $\mathbf{c}$. Further, $\tilde{\mathbf{s}}$ is a perturbed version of $\mathbf{s}$ (Assumption 2) and $\mathbf{m}_1$, $\mathbf{m}_2$ are modality-specific variables. The observations $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by two distinct mixing functions $\mathbf{f}_1 \neq \mathbf{f}_2$, which are applied to the subsets of latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$ respectively.

3.1 Latent variable model

On a high level, we assume that there exists a continuous random variable $\mathbf{z}$ that takes values in the latent space $\mathcal{Z} \subseteq \mathbb{R}^n$, which contains all information to generate observations of both modalities.[2] Moreover, we assume that $\mathbf{z} = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1, \mathbf{m}_2)$ can be uniquely partitioned into four disjoint parts:

(i) an invariant part $\mathbf{c}$ which is always shared across modalities, and which we refer to as *content*;

(ii) a variable part $\mathbf{s}$ which may change across modalities, and which we refer to as *style*;

(iii) two modality-specific parts, $\mathbf{m}_1$ and $\mathbf{m}_2$, each of which is unique to the respective modality.

Let $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 = (\tilde{\mathbf{c}}, \tilde{\mathbf{s}}, \mathbf{m}_2)$, where $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere and $\tilde{\mathbf{s}}$ is generated by perturbations that are specified in Section 3.2. Akin to multi-view ICA (Equation 1), we define the generative process for modalities $\mathbf{x}_1$ and $\mathbf{x}_2$ as follows:

$$\mathbf{z} \sim p_{\mathbf{z}}, \quad \mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}_1), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}_2), \tag{6}$$

where $\mathbf{f}_1 : \mathcal{Z}_1 \to \mathcal{X}_1$ and $\mathbf{f}_2 : \mathcal{Z}_2 \to \mathcal{X}_2$ are two smooth and invertible mixing functions with smooth inverse (i.e., diffeomorphisms) that generate observations $\mathbf{x}_1$ and $\mathbf{x}_2$ taking values in $\mathcal{X}_1 \subseteq \mathbb{R}^{d_1}$ and $\mathcal{X}_2 \subseteq \mathbb{R}^{d_2}$ respectively. Generally, we assume that observations from different modalities are generated by distinct mechanisms $\mathbf{f}_1 \neq \mathbf{f}_2$ that take modality-specific latent variables as input. As for the multi-view setting (von Kügelgen et al., 2021), the considered generative process *goes beyond the classical ICA setting* by allowing for statistical dependencies within blocks of variables (e.g., between dimensions of $\mathbf{c}$) and also for causal dependencies from content to style, as illustrated in Figure 1.

The outlined generative process is fairly general and it applies to a wide variety of practical settings. The content invariance describes a shared phenomenon that is not directly observed but manifests in the observations from both modalities. Style changes describe shared influences that are not robust across modalities, e.g., non-invertible transformations such as data augmentations, or non-deterministic effects of an unobserved confounder. Modality-specific factors can be viewed as variables that describe the inherent heterogeneity of each modality (e.g., background noise).

3.2 Relation between modalities

Next, we specify our assumptions on the relation between modalities by defining the conditional distribution $p_{\mathbf{z}_2|\mathbf{z}_1}$, which describes the relation between latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$, from which observations $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated via Equation (6). Similar to previous work in the multi-view setting (von Kügelgen et al., 2021), we assume that content is invariant, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 1), and that $\tilde{\mathbf{s}}$ is a perturbed version of $\mathbf{s}$ (Assumption 2). To state our assumptions for the multimodal setting, we also need to consider the modality-specific latent variables.

[2] Put differently, we assume that all observations lie on a continuous manifold, which can have much smaller dimensionality than the observation space of the respective modality.

Assumption 1 (Content-invariance)**.** *The conditional density $p_{\mathbf{z}_2|\mathbf{z}_1}$ over $\mathcal{Z}_2 \times \mathcal{Z}_1$ takes the form*

$$p_{\mathbf{z}_2|\mathbf{z}_1}(\mathbf{z}_2|\mathbf{z}_1) = \delta(\tilde{\mathbf{c}} - \mathbf{c}) p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) p_{\mathbf{m}_2}(\mathbf{m}_2) \tag{7}$$

for some continuous density $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ on $\mathcal{S} \times \mathcal{S}$, where $\delta(\cdot)$ is the Dirac delta function, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ a.e.

To fully specify $p_{\mathbf{z}_2|\mathbf{z}_1}$, it remains to define the style changes, which are described by the conditional distribution $p_{\tilde{\mathbf{s}}|\mathbf{s}}$. There are several justifications for modeling such a stochastic relation between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ (Zimmermann et al., 2021; von Kügelgen et al., 2021); one could either consider $\tilde{\mathbf{s}}$ to be a noisy version of $\mathbf{s}$, or consider $\tilde{\mathbf{s}}$ to be the result of an augmentation that induces a soft intervention on $\mathbf{s}$.[3]

Assumption 2 (Style changes)**.** *Let $\mathcal{A}$ be the powerset of style variables $\{1, \dots, n_s\}$ and let p_A be a distribution on $\mathcal{A}$. Then the style conditional $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ is obtained via conditioning on a set $A \sim p_A$:*

$$p_{\tilde{\mathbf{s}}|\mathbf{s},A}(\tilde{\mathbf{s}}|\mathbf{s}, A) = \delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A) \tag{8}$$

where $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a continuous density on $\mathcal{S}_A \times \mathcal{S}_A$, $\mathcal{S}_A \subseteq \mathcal{S}$ denotes the subspace of changing style variables specified by A, and $A^c = \{1, \dots, n_s\} \backslash A$ denotes the complement of A.

Further, for any style variable $l \in \{1, \dots, n_s\}$, there exists a set $A \subseteq \{1, \dots, n_s\}$ with $l \in A$, s.t.

(i) $p_A(A) > 0$,

(ii) $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ *is smooth w.r.t. both* $\tilde{\mathbf{s}}_A$, *and*

(iii) for any $\mathbf{s}_A$, $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$, *in some open non-empty subset containing* $\mathbf{s}_A$.

Intuitively, to generate a pair of observations $(\mathbf{x}_1, \mathbf{x}_2)$, we independently flip a biased coin for each style dimension to select a subset of style features $A \subseteq \{1, \dots, n_s\}$, which are jointly perturbed to obtain $\tilde{\mathbf{s}}$. Condition *(i)* ensures that every style dimension has a positive probability to be perturbed, while *(ii)* and *(iii)* are technical smoothness conditions that will be used for the proof of Theorem 1.

Summarizing, in this section we have formalized the multimodal generative process as a latent variable model (Section 3.1) and specified our assumptions on the relation between modalities via the conditional distribution $p_{\mathbf{z}_1|\mathbf{z}_2}$ (Section 3.2). Next, we segue into the topic of representation learning and show that, for the specified generative process, multimodal contrastive learning can identify the content factors up to a block-wise indeterminacy.

4 Identifiability Results

First, we need to define block-identifiability (von Kügelgen et al., 2021) for the multimodal setting in which we consider modality-specific mixing functions and encoders. In the following, n_c denotes the number of content variables and the subscript $1{:}n_c$ denotes the subset of content dimensions (indexed from 1 to n_c w.l.o.g.).

Definition 1 (Block-identifiability)**.** *The true content partition $\mathbf{c} = \mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c}$ is block-identified by a function $\mathbf{g}_i : \mathcal{X}_i \to \mathcal{Z}_i$, with $i \in \{1, 2\}$, if the inferred content partition $\hat{\mathbf{c}}_i = \mathbf{g}_i(\mathbf{x}_i)_{1:n_c}$ contains all and only information about $\mathbf{c}$, i.e., if there exists an invertible function $\mathbf{h}_i : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}}_i = \mathbf{h}_i(\mathbf{c})$.*

It is important to note that block-identifiability does not require the identification of *individual* factors, which is the goal of multi-view nonlinear ICA (Gresele et al., 2019; Locatello et al., 2020; Zimmermann et al., 2021; Klindt et al., 2021) and the basis for strict definitions of disentanglement (Bengio et al., 2013; Higgins et al., 2018; Shu et al., 2020). Instead, our goal is to isolate the group of invariant factors (i.e., the content partition) from the remaining factors of variation in the data.

Specifically, our goal is to show that contrastive learning can block-identify the content variables for the multimodal setting described in Section 3. We formalize this in Theorem 1 and thereby relax the assumptions from previous work by allowing for distinct generating mechanisms $\mathbf{f}_1 \neq \mathbf{f}_2$ and additional modality-specific latent variables.

[3]Note that the asymmetry between $\mathbf{z}_1$ and $\mathbf{z}_2$ (or between $\mathbf{s}$ and $\tilde{\mathbf{s}}$) is not strictly required. We chose to write $\mathbf{z}_2$ as a perturbation of $\mathbf{z}_1$ to simplify the notation and for consistency with previous work. Instead, we could model *both* $\mathbf{z}_1$ and $\mathbf{z}_2$ via pertubarbations of $\mathbf{z}$, as described in Appendix A.2.

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content $\mathbf{c}$	Style $\mathbf{s}$
1.0	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✗	✗	**0.99** $\pm$ 0.01	0.00 $\pm$ 0.00
0.75	✓	✗	**0.99** $\pm$ 0.00	0.52 $\pm$ 0.09
0.75	✗	✓	**1.00** $\pm$ 0.00	**0.79** $\pm$ 0.04
0.75	✓	✓	**0.99** $\pm$ 0.01	**0.81** $\pm$ 0.04

(a) Original setting

Generative process			R^2 (nonlinear)		
p(chg.)	Stat.	Cau.	Content $\mathbf{c}$	Style $\mathbf{s}$	Modality $\mathbf{m}_i$
1.0	✗	✗	**0.99** $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✓	✗	**0.95** $\pm$ 0.01	0.56 $\pm$ 0.23	0.00 $\pm$ 0.00
0.75	✗	✓	**0.98** $\pm$ 0.00	**0.87** $\pm$ 0.04	0.00 $\pm$ 0.00
0.75	✓	✓	**0.95** $\pm$ 0.03	**0.89** $\pm$ 0.07	0.00 $\pm$ 0.00

(b) Multimodal setting

Table 1: Results of the numerical simulations. We compare the original setting ($\mathbf{f}_1 = \mathbf{f}_2$, left table) with the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, right table). Only the multimodal setting includes modality-specific latent variables. Each row presents the results of a different setup with varying style-change probability p(chg.) and possible statistical (Stat.) and/or causal (Caus.) dependencies. Each value denotes the R^2 coefficient of determination (averaged across 3 seeds) for a nonlinear regression model that predicts the respective ground truth factor ($\mathbf{c}, \mathbf{s}$, or $\mathbf{m}_i$) from the learned representation.

Theorem 1. *Assume the data generating process described in Sec. 3.1, i.e. data pairs* $(\mathbf{x}_1, \mathbf{x}_2)$ *generated from Equation* (6) *with* $p_{\mathbf{z}_1} = p_{\mathbf{z}\setminus\{\mathbf{m}_2\}}$ *and* $p_{\mathbf{z}_2|\mathbf{z}_1}$ *as defined in Assumptions 1 and 2. Further, assume that* $p_{\mathbf{z}}$ *is a smooth and continuous density on* $\mathcal{Z}$ *with* $p_{\mathbf{z}}(\mathbf{z}) > 0$ *almost everywhere. Let* $\mathbf{g}_1 : \mathcal{X}_1 \to (0,1)^{n_c}$ *and* $\mathbf{g}_2 : \mathcal{X}_2 \to (0,1)^{n_c}$ *be smooth functions that minimize* $\mathcal{L}_{SymAlignMaxEnt}$ *as defined in Eq. (5). Then,* $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *block-identify the true content variables in the sense of Def. 1.*

A proof of Theorem 1 is provided in Appendix A.1. Intuitively, the result states that contrastive learning can identify the content variables up to a block-wise indeterminacy. Similar to previous work, the result is based on the optimization of the asymptotic form of the contrastive loss (Equation 4). Moreover, Theorem 1 assumes that the number of content variables is known or that it can be estimated (e.g., with a heuristic like the elbow method). We address the question of selecting the encoding size with dimensionality ablations throughout our experiments. In Section 7, we will return to the discussion of the assumptions in the context of the experimental results.

5 Experiments

The goal of our experiments is to test whether contrastive learning can block-identify content in the multimodal setting, as described by Theorem 1. First, we verify identifiability in a fully controlled setting with numerical simulations (Section 5.1). Second, we corroborate our findings on a complex multimodal dataset of image/text pairs (Section 5.2).

5.1 Numerical Simulation

We extend the numerical simulation from von Kügelgen et al. (2021) and implement the multimodal setting using modality-specific mixing functions ($\mathbf{f}_1 \neq \mathbf{f}_2$) with modality-specific latent variables. The numerical simulation allows us to measure identifiability with full control over the generative process. The data generation is consistent with the generative process described in Section 3. We sample $\mathbf{c} \sim \mathcal{N}(0, \Sigma_{\mathbf{c}})$, $\mathbf{m}_i \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_i})$, and $\mathbf{s} \sim \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_{\mathbf{s}})$. Statistical dependencies within blocks (e.g., among components of $\mathbf{c}$) are induced by non-zero off-diagonal entries in the corresponding covariance matrix (e.g. in $\Sigma_{\mathbf{c}}$). The causal dependence between content and style variables can be controlled via B. For style changes, Gaussian noise is added with probability π independently for each style dimension: $\tilde{\mathbf{s}}_i = \mathbf{s}_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$ with probability π. We generate the observations $\mathbf{x}_1 = \mathbf{f}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{x}_2 = \mathbf{f}_2(\mathbf{c}, \tilde{\mathbf{s}}, \mathbf{m}_2)$ using two *distinct* nonlinear mixing functions, i.e, for each $i \in \{1, 2\}$, $\mathbf{f}_i : \mathbb{R}^d \to \mathbb{R}^d$ is a separate, invertible 3-layer MLP with LeakyReLU($\alpha = 0.2$) activations. We train the encoders for 200,000 iterations using the symmetrized InfoNCE objective (Equation 4) and the hyperparameters listed in Appendix B.2. We evaluate block-identifiability by predicting the ground truth factors from the learned representation using kernel ridge regression and report the R^2 coefficient of determination on holdout data.

Results We compare the original setting ($\mathbf{f}_1 = \mathbf{f}_2$, Table 1a) with the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, Table 1b) and find that content can be block-identified in *both* settings, as the R^2 score is close to one for the prediction of content, and quasi-random for the prediction of style and modality-specific

information. Consistent with previous work, we observe that some style information can be predicted when there are statistical and/or causal dependencies; this is expected because statistical dependencies decrease the effective dimensionality of content, while the causal dependence $\mathbf{c} \to \mathbf{s}$ makes style partially predictable from the encoded content information. Overall, the results of the numerical simulation are consistent with our theoretical result from Theorem 1, showing that contrastive learning can block-identify content in the multimodal setting.

5.2 IMAGE/TEXT PAIRS

Next, we test whether block-identifiability holds in a more realistic setting with image/text pairs—two complex modalities with distinct generating mechanisms. We extend the *Causal3DIdent* dataset (von Kügelgen et al., 2021; Zimmermann et al., 2021), which allows us to measure and control the ground truth latent factors used to generate complex observations. We use *Blender* (Blender Online Community, 2018) to render high-dimensional images that depict a scene with a colored object illuminated by a differently colored spotlight and positioned in front of a colored background. The scene is defined by 11 latent factors: the shape of the object (7 classes), position of the object (x, y, z coordinates), orientation of the object (α, β, γ angles), position of the spotlight (θ angle), as well as the color of the object, background, and spotlight respectively (one numerical value for each).

Multimodal3DIdent We extend the *Causal3DIdent* dataset to the multimodal setting as follows. We generate textual descriptions from the latent factors by adapting the text rendering from the *CLEVR* dataset (Johnson et al., 2017). Each image/text pair shares information about the shape of the object (cow, teapot, etc.) and its position in the scene (e.g., bottom-right). For each position factor, we use three clearly discernable values (top/center/bottom; left/center/right), which can be described in text more naturally than coordinates. While shape and position are always shared (i.e., content) between the paired image and text, the color of the object is causally influenced by position (cf. Appendix B.1) and is stochastically shared (i.e., style). For the object color, we use a continuous hue value, whereas for the text we match the RGB value with the nearest value from a given palette (i.e., a list of named colors, such as brown, beige, olive, etc.). The color palette is randomly sampled from a set of three palettes[4] to ensure the object color depicted in the image does not uniquely determine the color described in the text. As modality-specific factors for the images, we have object rotation, spotlight position, and background color, while for the textual descriptions, we follow Johnson et al. (2017) and use 5 different types of phrases to introduce modality-specific variation. Examples of image/text pairs are shown in Figure 2.

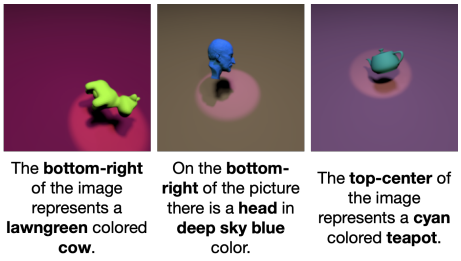
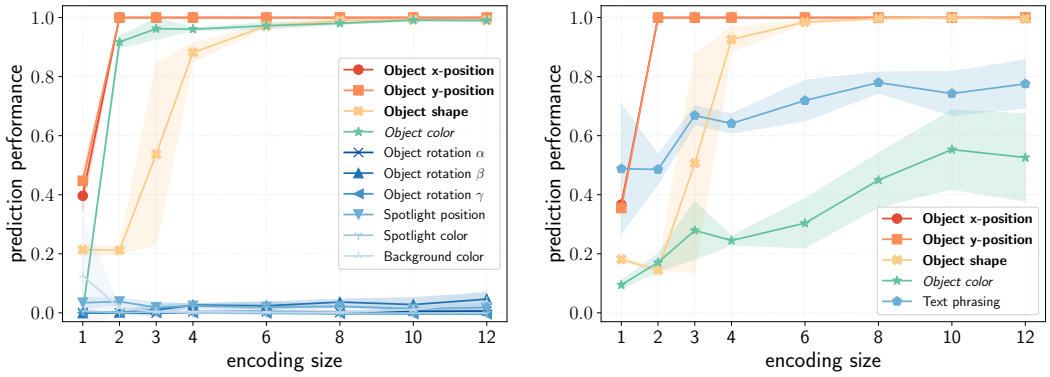


Figure 2: Examples of image/text pairs.

We train the encoders for 100,000 iterations using the symmetrized InfoNCE objective (Equation 4) and the hyperparameters listed in Appendix B.2. For the image encoder we use a ResNet-18 architecture (He et al., 2016) and for the text we use 1D convolutions. As for the numerical simulation, we evaluate block-identifiability by predicting the ground truth factors from the learned representation. For continuous factors, we use kernel ridge regression and report the R^2 score, whereas for discrete factors we report the classification accuracy of an MLP with a single hidden layer.

Results Figure 3 presents the results on *Multimodal3DIdent* with a dimensionality ablation, where we vary the size of the encoding of the model. Content factors (object position and shape) are always encoded well, unless the encoding size is too small (i.e. smaller than four). When there is sufficient capacity, style information (object color) is also encoded, partly because there is a causal dependence from content to style and partly because of the excess capacity, as already observed in previous work. Image-specific information (object rotation, spotlight position, background color) is mostly discarded, independent of the encoding size. Text-specific information (phrasing) is encoded to a moderate degree (48–80% accuracy), which we attribute to the fact that phrasing is a discrete factor that violates the assumption of continuous latents. This hints at possible limitations in the presence

[4] We use the following three palettes from the `matplotlib.colors` API: Tableau colors (10 values), CSS4 colors (148 values), and XKCD colors (949 values).



(a) Prediction of image factors

(b) Prediction of text factors

Figure 3: Result on *Multimodal3DIdent* as a function of the encoding size of the model. We assess the nonlinear prediction of ground-truth image factors (left subplot) and text factors (right subplot) to quantify how well the learned representation encodes the respective factors. Content factors are denoted in bold and style factors in italic. Along the x-axis, we vary the encoding size, i.e., the output dimensionality of the model. We measure the prediction performance in terms of the R^2 coefficient of determination for continuous factors and classification accuracy for discrete factors respectively. Each point denotes the average across three seeds and bands show one standard deviation.

of discrete latent factors, which we further investigate in Appendix B.2 and discuss in Section 7. Overall, our results suggest that contrastive learning can block-identify content factors in a complex multimodal setting with image/text pairs.

6 Related work

Multi-view nonlinear ICA The goal of multi-view nonlinear ICA is to identify latent factors shared between different views, as described in Section 2.1. There is a thread of works (Gresele et al., 2019; Lyu and Fu, 2020; Locatello et al., 2020) that recover the latent variable up to a component-wise indeterminacy in a setting with mutually independent latent components. Beyond the classical assumption of independent components, there is a line of works (von Kügelgen et al., 2021; Kong et al., 2022) that partition the latent space into blocks of invariant and blocks of changing components and show that the invariant components can be identified up to a block-wise indeterminacy, even when there are nontrivial dependencies between latent components. Our work advances in this direction and considers heterogenous modalities with nontrivial statistical and causal dependencies between latents. We prove that shared factors can be block-identified in a novel setting with modality-specific mixing functions and modality-specific latent variables.

Multimodal representation learning Multimodal representation learning seeks to integrate information from heterogeneous sources into a joint representation (Baltrušaitis et al., 2019; Guo et al., 2019). There is a myriad of methods designed to learn representations of multimodal data either directly or indirectly. Among methods that learn representations indirectly, there are multimodal autoencoders (Ngiam et al., 2011; Geng et al., 2022; Bachmann et al., 2022; Aghajanyan et al., 2022) and a large variety of multimodal generative models (e.g., Suzuki et al., 2016; Wu and Goodman, 2018; Shi et al., 2019; Huang et al., 2018; Tsai et al., 2019; Ramesh et al., 2021) that learn representations by backpropagation of different forms of reconstruction and/or masked prediction error. A more direct approach is taken by decoder-free methods that maximize the similarity between the encodings of different modalities. This class of methods includes nonlinear canonical correlation analysis (Akaho, 2001; Bach and Jordan, 2002; Andrew et al., 2013; Wang et al., 2016; Tang et al., 2017; Karami and Schuurmans, 2021) as well as multi-view and multimodal contrastive learning (Tian et al., 2019; Tsai et al., 2021; Federici et al., 2020; Radford et al., 2021; Poklukar et al., 2022). While all of the named methods aim to integrate information across modalities, they do not answer the underlying question of identifiability, which our work seeks to address.

7 DISCUSSION

Implications and scope We have shown that contrastive learning can block-identify shared factors in the multimodal setting. Numerical simulations (Section 5.1) verify our main theoretical result (Theorem 1), showing that contrastive learning encodes *all and only* content information (Definition 1), when the size of the encoding matches the number of content factors. Experiments on a complex dataset of image/text pairs corroborate that contrastive learning can isolate content in a more realistic setting and even under some violations of the assumptions underlying Theorem 1. More generally, we observe that contrastive learning encodes invariant information (i.e., content) very well across all settings. When there is sufficient capacity, stochastically shared information (i.e., style) is encoded to a moderate degree, but without affecting the prediction of invariant information. For practice, our results suggests that contrastive learning without capacity constraints can encode *any* shared factor, irregardless of whether the factor is truly invariant across modalities or if its effect on the observations is confounded by noise or other factors. This is in line with the information-theoretic view (Oord et al., 2018; Poole et al., 2019), i.e., that contrastive learning maximizes the mutual information between representations—a measure of mutual dependence that quantifies *any* information that is shared. Our results demonstrate that the size of the encoding can be reduced to encode *all and only* invariant information, as captured by the notion of block-identifiability. In practice, this can be leveraged for representation learning in settings of content-preserving distribution shifts (Mitrovic et al., 2021; Federici et al., 2021), where information relevant for a downstream task remains unchanged.

Limitations and outlook First, Theorem 1 assumes that only *invariant* factors can be block-identified. However, in practice, there can be pairs of observations for which the invariance is inadvertently violated, e.g., due to measurement errors, occlusions, or other mistakes in the data collection. On the one hand, such a violation can be viewed as a mere artifact of the data collection and could be managed via interventions on the generative process, e.g., actions in reinforcement learning (Lippe et al., 2022; Brehmer et al., 2022; Ahuja et al., 2022). On the other hand, violations of the content-invariance blur the line between content and style factors and it would be interesting to study identifiability in a more general setting with *only* stochastically shared factors. Second, Theorem 1 assumes that the number of content factors is known or that it can be estimated. In practice, this might not be a significant limitation, since the number of content factors can be viewed as a single hyperparameter (e.g., Locatello et al., 2020), though the design of suitable heuristics is an interesting research direction. Third, Theorem 1 assumes that all latent factors are continuous. While this assumption prevails in related work (Hyvärinen and Pajunen, 1999; Hyvarinen and Morioka, 2016; Hyvärinen et al., 2019; Gresele et al., 2019; Locatello et al., 2019; 2020; Zimmermann et al., 2021; von Kügelgen et al., 2021; Klindt et al., 2021), our results in Figure 3b indicate that in the presence of discrete factors some style or modality-specific information can be encoded. In Appendix B.2, we provide a simple numerical simulation that supports these findings. Finally, our model can be extended to more than two modalities—a setting for which there are intriguing identifiability results (Gresele et al., 2019; Schölkopf et al., 2016). Summarizing, the described limitations mirror the assumptions of the generative process (Section 3), which may be relaxed in future work.

8 CONCLUSION

We addressed the problem of identifiability for multimodal representation learning and showed that contrastive learning can block-identify latent factors shared between heterogenous modalities. We formalize the multimodal generative process as a novel latent variable model with modality-specific generative mechanisms and nontrivial statistical and causal dependencies between latents. We prove that contrastive learning can identify shared latent factors up to a block-wise indeterminacy and therefore isolate invariances between modalities from other changeable factors. Our theoretical results are corroborated by numerical simulations and on a complex multimodal dataset of image/text pairs. More generally, we believe that our work will help in shaping a theoretical foundation for multimodal representation learning and that further relaxations of the presented generative process offer rich opportunities for future work.

REPRODUCIBILITY STATEMENT

For our theoretical statements, we provide detailed derivations and state the necessary assumptions. The generative process is specified in Section 3 and the assumptions for block-identifiability are referenced in Theorem 1. We test violations of the key assumptions with suitable experiments (dimensionality ablation, discrete latent factors) and discuss the limitations of our work in Section 7. Further, we empirically verify our theoretical results with numerical simulations and on complex multimodal data. To ensure empirical reproducibility, the results of every experiment were averaged over multiple seeds and are reported with standard deviations. Information about implementation details, hyperparameter settings, and evaluation metrics are included in Appendix B.1. In the supplementary material, we provide the code to reproduce the experiments.

REFERENCES

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., and Zettlemoyer, L. (2022). CM3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

Ahuja, K., Hartford, J., and Bengio, Y. (2022). Weakly supervised representation learning with sparse perturbations. *arXiv preprint arXiv:2206.01101*.

Akaho, S. (2001). A kernel method for canonical correlation analysis. *arXiv preprint arxiv:cs/0609071*.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning*.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3:1–48.

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal Machine Learning: A survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Blender Online Community (2018). *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. (2022). Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*.

Chen, J. and Batmanghelich, K. (2020). Weakly supervised disentanglement by pairwise similarities. In *AAAI Conference on Artificial Intelligence*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*.

Darmois, G. (1951). Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. (2020). Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*.

Federici, M., Tomioka, R., and Forré, P. (2021). An information-theoretic approach to distribution shifts. In *Advances in Neural Information Processing Systems*.

Geng, X., Liu, H., Lee, L., Schuurams, D., Levine, S., and Abbeel, P. (2022). Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2019). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Conference on Uncertainty in Artificial Intelligence*.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*.

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

Huang, X., Liu, M., Belongie, S. J., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29.

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439.

Hyvärinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition*.

Karami, M. and Schuurmans, D. (2021). Deep probabilistic canonical correlation analysis. In *AAAI Conference on Artificial Intelligence*.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020). Variational autoencoders and nonlinear ICA: a unifying framework. In *International Conference on Artificial Intelligence and Statistics*.

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. M. (2021). Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*.

Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., and Zhang, K. (2022). Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lippe, P., Sarah, M., S., L., M., A. Y., Taco, C., and S., G. (2022). Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*.

Lyu, Q. and Fu, X. (2020). Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Trans. Signal Process.*, 68:2697–2712.

Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. (2021). Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *International Conference on Machine Learning*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Poklukar, P., Vasco, M., Yin, H., Melo, F. S., Paiva, A., and Kragic, D. (2022). Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Salakhutdinov, R. and Hinton, G. E. (2009). Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*.

Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. (2016). Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398.

Shi, Y., Siddharth, N., Paige, B., and Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. (2020). Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*.

Song, L., Anandkumar, A., Dai, B., and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Tang, Q., Wang, W., and Livescu, K. (2017). Acoustic feature learning via deep variational canonical correlation analysis. In *INTERSPEECH*.

Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Conference of the Association for Computational Linguistics*.

Tsai, Y. H., Wu, Y., Salakhutdinov, R., and Morency, L. (2021). Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*.

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*.

Wang, W., Lee, H., and Livescu, K. (2016). Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*.

Wu, M. and Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. (2020). Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

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

A THEORY

A.1 PROOF OF THEOREM 1

Theorem 1. *Assume the data generating process described in Sec. 3.1, i.e. data pairs* $(\mathbf{x}_1, \mathbf{x}_2)$ *generated from Equation (6) with* $p_{\mathbf{z}_1} = p_{\mathbf{z}\setminus\{\mathbf{m}_2\}}$ *and* $p_{\mathbf{z}_2|\mathbf{z}_1}$ *as defined in Assumptions 1 and 2. Further, assume that* $p_{\mathbf{z}}$ *is a smooth and continuous density on* $\mathcal{Z}$ *with* $p_{\mathbf{z}}(\mathbf{z}) > 0$ *almost everywhere. Let* $\mathbf{g}_1 : \mathcal{X}_1 \to (0,1)^{n_c}$ *and* $\mathbf{g}_2 : \mathcal{X}_2 \to (0,1)^{n_c}$ *be smooth functions that minimize* $\mathcal{L}_{\textit{SymAlignMaxEnt}}$ *as defined in Eq. (5). Then,* $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *block-identify the true content variables in the sense of Def. 1.*

Proof. To prove Theorem 1, we follow the proof structure from von Kügelgen et al. (2021, Theorem 4.4) and divide the proof into three steps. First, we show that there exists a pair of smooth functions $\mathbf{g}_1^*, \mathbf{g}_2^*$ that attain the global minimum of $\mathcal{L}_{\text{SymAlignMaxEnt}}$ (Eq. 5). Further, in Equations (13–15), we derive invariance conditions that have to hold almost surely for any pair of smooth functions $\mathbf{g}_1, \mathbf{g}_2$ attaining the global minimum of Eq. (5). In Step 2, we use the invariance conditions derived in Step 1 to show by contradiction that *any* pair of smooth functions $\mathbf{g}_1, \mathbf{g}_2$ that attain the global minimum in Eq. (5) can only depend on content and not on style or modality-specific information. In the third and final step, for $\mathbf{h}_1 := \mathbf{g}_1 \circ \mathbf{f}_1$ and $\mathbf{h}_2 := \mathbf{g}_2 \circ \mathbf{f}_2$, we show that both functions must be bijections and hence that $\mathbf{c}$ is block-identified by $\mathbf{g}_1$ and $\mathbf{g}_2$ respectively.

Step 1. Recall the asymptotic form of the objective, as defined in Equation (5):

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}} \left[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2\right] - {}^1\!/_2 \left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right) \ . \tag{5}$$

The global minimum of $\mathcal{L}_{\text{SymAlignMaxEnt}}$ is reached when the first term is minimized and the second term is maximized. The first term is minimized when the encoders $\mathbf{g}_1$ and $\mathbf{g}_2$ are perfectly aligned, i.e., when $\mathbf{g}_1(\mathbf{x}_1) = \mathbf{g}_2(\mathbf{x}_2)$ holds for all pairs $(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1,\mathbf{x}_2}$. The second term attains its maximum when $\mathbf{g}_1$ and $\mathbf{g}_2$ map to a uniformly distributed random variable on $(0,1)^{n_c}$ respectively.[5]

To show that there *exists* a pair of functions that minimize $\mathcal{L}_{\text{SymAlignMaxEnt}}$, let $\mathbf{g}_1^* := \boldsymbol{d}_1 \circ \mathbf{f}_{1,1:n_c}^{-1}$ and let $\mathbf{g}_2^* := \boldsymbol{d}_2 \circ \mathbf{f}_{2,1:n_c}^{-1}$, where the subscript $1{:}n_c$ indexes the subset of content dimensions w.l.o.g. and where $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ will be defined using the Darmois construction (Darmois, 1951; Hyvärinen and Pajunen, 1999). First, recall that $\mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{c}$ and that $\mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c} = \tilde{\mathbf{c}}$ by definition. Second, for $i \in \{1, 2\}$, let us define $\boldsymbol{d}_i : \mathcal{C} \mapsto (0,1)^{n_c}$ using the Darmois construction, such that $\boldsymbol{d}_i$ maps $\mathbf{c}$ and $\tilde{\mathbf{c}}$ to a uniform random variable respectively. It follows that $\mathbf{g}_1^*, \mathbf{g}_2^*$ are smooth functions, because any function $\boldsymbol{d}_i$ obtained via the Darmois construction is smooth and $\mathbf{f}_1^{-1}, \mathbf{f}_2^{-1}$ are smooth as well (each being the inverse of a smooth function).

Next, we show that the pair of functions $\mathbf{g}_1^*, \mathbf{g}_2^*$, as defined above, attains the global minimum of the objective $\mathcal{L}_{\text{SymAlignMaxEnt}}$. We have that

$$\begin{aligned}
\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1^*, \mathbf{g}_2^*) &= \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}} \left[\|\mathbf{g}_1^*(\mathbf{x}_1) - \mathbf{g}_2^*(\mathbf{x}_2)\|_2\right] - {}^1\!/_2 \left(H(\mathbf{g}_1^*(\mathbf{x}_1)) + H(\mathbf{g}_2^*(\mathbf{x}_2))\right) && (9)\\
&= \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}} \left[\|\boldsymbol{d}_1(\mathbf{c}) - \boldsymbol{d}_2(\tilde{\mathbf{c}})\|_2\right] - {}^1\!/_2 \left(H(\boldsymbol{d}_1(\mathbf{c})) + H(\boldsymbol{d}_2(\tilde{\mathbf{c}}))\right) && (10)\\
&= 0 && (11)
\end{aligned}$$

where by Assumption 1, $\mathbf{c} = \tilde{\mathbf{c}}$ almost surely, which implies that the first term is zero almost surely. Further, $\boldsymbol{d}_i$ maps $\mathbf{c}, \tilde{\mathbf{c}}$ to uniformly distributed random variables on $(0,1)^{n_c}$, which implies that the differential entropy of $\boldsymbol{d}_1(\mathbf{c})$ and $\boldsymbol{d}_2(\tilde{\mathbf{c}})$ is zero, as well. Consequently, there exists a pair of functions $\mathbf{g}_1^*, \mathbf{g}_2^*$ that minimizes $\mathcal{L}_{\text{SymAlignMaxEnt}}$.

Next, let $\mathbf{g}_1 : \mathcal{X}_1 \mapsto (0,1)^{n_c}$ and $\mathbf{g}_2 : \mathcal{X}_2 \mapsto (0,1)^{n_c}$ be *any* pair of smooth functions that attains the global minimum of Eq. (5), i.e.,

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}} \left[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2\right] - {}^1\!/_2 \left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right) = 0 \ . \tag{12}$$

Let $\mathbf{h}_1 := \mathbf{g}_1 \circ \mathbf{f}_1$ and $\mathbf{h}_2 := \mathbf{g}_2 \circ \mathbf{f}_2$, and notice that both are smooth functions since all involved functions are smooth by definition. Since Equation (12) is a global minimum, it implies the following

[5] Note that we restrict the range of $\mathbf{g}_1$ and $\mathbf{g}_2$ to $(0,1)^{n_c}$ by definition merely to simplify the notation. Generally, the uniform distribution $\mathcal{U}(a, b)$ is the maximum entropy distribution on the interval $[a, b]$.

invariance conditions for the individual terms:

$$\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\|\mathbf{h}_1(\mathbf{z}_1)-\mathbf{h}_2(\mathbf{z}_2)\|_2\right]=0 \tag{13}$$

$$H(\mathbf{h}_1(\mathbf{z}_2))=0 \tag{14}$$

$$H(\mathbf{h}_2(\mathbf{z}_2))=0 \tag{15}$$

Hence, $\mathbf{h}_1(\mathbf{z}_1)=\mathbf{h}_2(\mathbf{z}_2)$ must hold almost surely w.r.t. $p_{\mathbf{x}_1,\mathbf{x}_2}$. Additionally, Equation (14) (resp. Equation (15)) implies that $\hat{\mathbf{c}}_1=\mathbf{h}_1(\mathbf{z}_1)$ (resp. $\hat{\mathbf{c}}_2=\mathbf{h}_2(\mathbf{z}_2)$) must be uniform on $(0,1)^{n_c}$.

Step 2. Next, we show that any pair of functions that minimize $\mathcal{L}_{\text{SymAlignMaxEnt}}$ depend only on content information. Since style is independent of $\mathbf{m}_1$ and $\mathbf{m}_2$, we first show that $\mathbf{h}_1(\mathbf{z}_1)$ does not depend on $\mathbf{m}_1$, and that $\mathbf{h}_2(\mathbf{z}_2)$ does not depend on $\mathbf{m}_2$. We then show that $\mathbf{h}_1$ and $\mathbf{h}_2$ also cannot depend on style, based on a result from previous work.

First note, that we can exclude all degenerate solutions where $\mathbf{g}_1$ maps a component of $\mathbf{m}_1$ to a constant, since $\mathbf{g}_1$ would not be invertible anymore and such a solution would violate the invariance in Eq. (14). To prove a contradiction, suppose that, w.l.o.g., $\mathbf{h}_1(\mathbf{c},\mathbf{s},\mathbf{m}_1)_{1:n_c}:=\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ depends on some component in $\mathbf{m}_1$. By Assumption 1, $\mathbf{z}_2$ is independent of $\mathbf{m}_1$ and hence

$$|(\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}-\mathbf{h}_2(\mathbf{z}_2)_{1:n_c})|>0 \tag{16}$$

holds almost surely if $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ depends on any component of $\mathbf{m}_1$, which violates the invariance in Equation (13). Thus, $\mathbf{h}_1(\mathbf{z}_1)_{1:n_\mathbf{c}}$ cannot depend on any component of $\mathbf{m}_1$.

Having established that neither $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$, nor $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ can depend on modality-specific information, it remains to show that style information is not encoded, as well. Leveraging Assumption 2, we can show that the strict inequality in Equation (13) has a positve density if $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ or $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ was dependent on a dimension in $\mathbf{s}$ respectively $\tilde{\mathbf{s}}$, which would again lead to a violation of the invariance derived in Equation (13), as shown in von Kügelgen et al. (2021, Proof of Theorem 4.2).

Step 3. It remains to show that $\mathbf{h}_1,\mathbf{h}_2$ are bijections. We know that $\mathcal{C}$ and $(0,1)^{n_c}$ are simply connected and oriented C^1 manifolds, and we have established in Step 1 that $\mathbf{h}_1$ and $\mathbf{h}_2$ are smooth and hence differentiable functions. Since $p_\mathbf{c}$ is a regular density, the uniform distributions w.r.t. the pushthrough functions $\mathbf{h}_1$ and $\mathbf{h}_2$ are regular densities. Thus, $\mathbf{h}_1$ and $\mathbf{h}_2$ are bijections (Zimmermann et al., 2021, Proposition 5)

Step 3 concludes the proof. We have shown that for any pair of smooth functions $\mathbf{g}_1,\mathbf{g}_2$ that attain the global minimum of Eq. (5), we have that $\mathbf{c}$ is *block-identified* (Def. 1) by $\mathbf{g}_1$ and $\mathbf{g}_2$. □

A.2 Symmetric generative process

Throughout the main body of the paper, we described an asymmetric generating mechanism, where $\mathbf{z}_2$ is a perturbed version of $\mathbf{z}_1$. Here, we will briefly sketch out how our model and results can be adapted to a symmetric setting, where *both* $\mathbf{z}_1$ and $\mathbf{z}_2$ are generated as perturbations of $\mathbf{z}$.

Concretely, we would need to make small adjustments to Assumptions 1 and 2 as follows. We start with the content invariance in Assumption 1, which specifies how $\mathbf{z}_1=(\tilde{\mathbf{c}}_1,\tilde{\mathbf{s}}_1,\tilde{\mathbf{m}}_1)$ and $\mathbf{z}_2=(\tilde{\mathbf{c}}_2,\tilde{\mathbf{s}}_2,\tilde{\mathbf{m}}_2)$ are generated.

Let $i\in\{1,2\}$. The conditional density $p_{\mathbf{z}_i|\mathbf{z}}$ over $\mathcal{Z}_i\times\mathcal{Z}$ takes the form

$$p_{\mathbf{z}_i|\mathbf{z}}(\mathbf{z}_i|\mathbf{z})=\delta(\tilde{\mathbf{c}}_i-\mathbf{c})\delta(\tilde{\mathbf{m}}_i-\mathbf{m}_i)p_{\tilde{\mathbf{s}}_i|\mathbf{s}}(\tilde{\mathbf{s}}_i|\mathbf{s})\ , \tag{17}$$

where $\delta(\cdot)$ is the Dirac delta function, i.e., $\tilde{\mathbf{c}}_i=\mathbf{c}$ almost everywhere, as well as $\tilde{\mathbf{m}}_i=\mathbf{m}_i$ almost everywhere. Note that since $\tilde{\mathbf{c}}_1=\mathbf{c}$ a.e. and $\mathbf{c}=\tilde{\mathbf{c}}_2$ a.e., it follows that $\tilde{\mathbf{c}}_1=\tilde{\mathbf{c}}_2$ almost everywhere, which is a property that is needed in Step 1 in the proof of Theorem 1. In addition, it still holds that $\tilde{\mathbf{m}}_i\perp\!\!\!\perp\mathbf{z}_j$, for $i,j\in\{1,2\}$ and $i\neq j$, which is needed in Step 2 of the proof to show that modality-specific information is not encoded.

Lastly, we need to revisit Assumption 2, for which both $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ would be generated through perturbations of $\mathbf{s}$ via the conditional distribution $p_{\tilde{\mathbf{s}}_i|\mathbf{s}}$ on $\mathcal{S}\times\mathcal{S}$, as described in Assumption 2, for each i individually. As a small technical nuance, we would need to specify the conditional generation of the perturbed style variables $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ such that they are not perturbed in an identical manner w.r.t. $\mathbf{s}$. This can be ensured by, e.g., constraining p_A appropriately to exclude the degenerate case where dimensions in $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ are perfectly aligned—a case that needs to be excluded for Step 2 of the proof of Theorem 1.

B EXPERIMENTS

B.1 EXPERIMENTAL DETAILS

Numerical simulation The generative process is described in Section 5.1. Here, we provide additional information about the experiment. The invertible MLP is constructed similar to previous work (Hyvarinen and Morioka, 2016; Hyvärinen and Morioka, 2017; Zimmermann et al., 2021; von Kügelgen et al., 2021) by resampling square weight matrices until their condition number surpasses a threshold value. For the original setting ($\mathbf{f}_1 = \mathbf{f}_2$), we use one encoder ($\mathbf{g}_1 = \mathbf{g}_2$), whereas for the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$), we use distinct encoders ($\mathbf{g}_1 \neq \mathbf{g}_2$) to mirror the assumption of distinct mixing functions and because, in practice, the dimensionality of the observations can differ across modalities.

Multimodal3DIdent Our dataset of image/text pairs is based on the code used to generate the *Causal3DIdent* (von Kügelgen et al., 2021; Zimmermann et al., 2021) and *CLEVR* (Johnson et al., 2017) datasets. Images are generated using the *Blender* renderer (Blender Online Community, 2018). The rendering serves as a complex mixing function that generates the images from 11 different parameters (i.e., latent factors) that are listed in Table 3. To generate textual descriptions, we adapt the text rendering from *CLEVR* (Johnson et al., 2017) and use 5 different phrases to induce modality-specific variation. The latent factors used to generate the text are also listed in Table 3. The dependence between the image and text modality is determined by three content factors (object shape, x-position, and y-position) and one style factor (object color). For the object color in the image, we use a continuous hue value, whereas for the text we match the RGB value with the nearest color value from one of three different palettes[4] that is sampled uniformly at random for each observation. Further, we ensure that there are no overlapping color values across palettes by using a prefix for the respective palette (e.g., "xkcd:black") when necessary. In Section 5.2, we use a version of the *Multimodal3DIdent* dataset with a causal dependence from content to style. Specifically, the color of the object depends on its x-position. In particular, we split the range of hue values $[0, 2\pi]$ into three equally sized intervals and associate each of these intervals with a fixed x-position of the object. For instance, if x-position is "left", we sample the hue value from the interval $[0, {}^2\!/_3\,\pi]$. Consequently, the color of the object can be predicted to some degree from the position of the object. Samples of image/text pairs from the *Multimodal3DIdent* dataset are shown in Figures 2 and 4. The hyperparameters for the experiment are listed in Table 2b.

Parameter	**Value**
Generating function	3-layer MLP
Encoder	7-layer MLP
Optimizer	Adam
Cond. threshold ratio	1e-3
Dimensionality d	15
Batch size	6144
Learning rate	1e-4
Temperature τ	1.0
# Seeds	3
# Iterations	300,000
Similarity metric	Euclidian
Gradient clipping	2-norm; max value 2

(a) Parameters used for the numerical simulation.

Parameter	**Value**
Generating function	Image and text rendering
Image encoder	ResNet-18
Text encoder	4-layer ConvNet
Optimizer	Adam
Batch size	256
Learning rate	1e-5
Temperature τ	1.0
# Seeds	3
# Iterations	100,000
# Samples (train / val / test)	125,000 / 10,000 / 10,000
Similarity metric	Cosine similarity
Gradient clipping	2-norm; max value 1

(b) Parameters used for *Multimodal3DIdent*.

Table 2: Experimental parameters and hyperparameters used for the two experiments in the main text.

On the **bottom-left** of the image there is a **tab:cyan** object in the form of a **dragon.** | On the **top-center** of the image there is a **tab:cyan** object in the form of a **head**. | A **green yellow teapot** is on the **mid-right** of the image. | An **horse** of **tab:pink color** is visible positioned in the **bottom-center** of the image. | A **lawn green cow** is on the **bottom-right** of the image. | A **xkcd:poison green hare** is on the **top-left** of the image. | A **spring green armadillo** is on the **bottom-left** of the image.

Figure 4: Examples of image/text pairs from the *Multimodal3DIdent* dataset. Each sample shows one of the 7 shapes (i.e., classes) of objects included in the dataset.

Latent factor	**Distribution**	**Details**
Object shape	Categorical	7 unique values, drawn uniformly
Object x-position	Categorical	3 unique values, drawn uniformly
Object y-position	Categorical	3 unique values, drawn uniformly
Object color	Uniform	hue value in $[0, 2\pi]$
Object rotation α	Uniform	angle in $[0, 2\pi]$
Object rotation β	Uniform	angle in $[0, 2\pi]$
Object rotation γ	Uniform	angle in $[0, 2\pi]$
Spotlight position	Uniform	angle in $[0, 2\pi]$
Spotlight color	Uniform	hue value in $[0, 2\pi]$
Background color	Uniform	hue value in $[0, 2\pi]$
Object shape	Categorical	7 unique values, drawn uniformly
Object x-position	Categorical	3 unique values, drawn uniformly
Object y-position	Categorical	3 unique values, drawn uniformly
Object color	Categorical	drawn uniformly from 3 palettes[4]
Text phrasing	Categorical	5 unique values, drawn uniformly

Table 3: Description of the latent factors used to generate *Multimodal3DIdent*. The first 11 factors are used to generate the images and the remaining 5 factors are used to generate the text. Content factors are denoted in bold and style factors in italic. The remaining factors are modality-specific.

B.2 Additional experimental results

Numerical simulation without modality-specific latents Recall that the considered generative process (Section 3) has two sources of modality-specific variation: modality-specific mixing functions and modality-specific latent variables. To decouple the effect of these two sources of variation, we conduct an ablation study *without* modality-specific latent variables. Table 4 presents the results, showing that content is block-identified in both the original setting ($\mathbf{f}_1 = \mathbf{f}_2$, Table 4a) and the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, Table 4b). Compared to Table 1, we observe that the content prediction is improved slightly in the case without modality-specific latent variables. Hence, our results suggest that contrastive learning can block-identify content factors in the multimodal setting with and without modality-specific latent variables.

Numerical simulation with discrete style Extending the numerical simulation from von Kügelgen et al. (2021), we test block-identifiability of content information when observations are defined by a mixture of continuous content and discrete style variables. In this setting, content is a *continuous* random variable with 5 components sampled from a uniform distribution, while style is a *discrete* variable with 5 components sampled from a multinomial distribution with 10 classes. We train an encoder with the InfoNCE objective and set the encoding size to 5 dimensions. Our results in Table 5 suggest that style information is encoded to a small degree ($R^2 = 0.23$) above chance level ($R^2 = 0.2$) when style is discrete.

	R^2 (nonlinear)	
Style factors	**content c**	**style s**
Continuous	$\mathbf{0.95} \pm 0.04$	0.00 ± 0.01
Discrete	$\mathbf{0.94} \pm 0.02$	0.23 ± 0.01

Table 5: Numerical simulations comparing continuous vs. discrete style factors.

Generative process			R^2 (nonlinear)	
p(chg.)	**Stat.**	**Cau.**	**Content c**	**Style s**
1.0	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✗	✗	**0.99** $\pm$ 0.01	0.00 $\pm$ 0.00
0.75	✓	✗	**0.99** $\pm$ 0.00	0.52 $\pm$ 0.09
0.75	✗	✓	**1.00** $\pm$ 0.00	**0.79** $\pm$ 0.04
0.75	✓	✓	**0.99** $\pm$ 0.01	**0.81** $\pm$ 0.04

(a) Original setting

Generative process			R^2 (nonlinear)	
p(chg.)	**Stat.**	**Cau.**	**Content c**	**Style s**
1.0	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✓	✗	**0.99** $\pm$ 0.01	0.36 $\pm$ 0.10
0.75	✗	✓	**1.00** $\pm$ 0.00	**0.81** $\pm$ 0.03
0.75	✓	✓	**0.99** $\pm$ 0.01	**0.83** $\pm$ 0.05

(b) Multimodal setting

Table 4: Results of the numerical simulations *without* modality-specific latent variables. We compare the original setting ($\mathbf{f}_1 = \mathbf{f}_2$, left table) with the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, right table). Each row presents the results of a different setup with varying style-change probability p(chg.) and possible statistical (Stat.) and/or causal (Caus.) dependencies. Each value denotes the R^2 coefficient of determination (averaged across 3 seeds) for a nonlinear regression model that predicts the respective ground truth factor ($\mathbf{c}, \mathbf{s}$, or $\mathbf{m}_i$) from the learned representation.