IDENTIFIABILITY RESULTS FOR MULTIMODAL CONTRASTIVE LEARNING

多模态对比学习的可识别性结果

Imant Daunhawer [1,†] , Alice Bizeul [1,2] , Emanuele Palumbo [1,2] , Alexander Marx [1,2,*] & Julia E. Vogt [1,*]

Imant Daunhawer [1,†] , Alice Bizeul [1,2] , Emanuele Palumbo [1,2] , Alexander Marx [1,2,*] & Julia E. Vogt [1,*]

[1] Department of Computer Science, ETH Zurich
[1] 苏黎世联邦理工学院计算机科学系
[2] ETH AI Center, ETH Zurich
[2] 苏黎世联邦理工学院人工智能中心

ABSTRACT

摘要

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.

对比学习是近年来多视角和多模态学习进展的基石，例如，在图像/标题对的表示学习中。尽管其有效性尚未完全理解，但最近的一系列研究表明，对比学习可以逆转数据生成过程，并恢复视角之间共享的真实潜在因素。在本研究中，我们提出了多模态对比学习的新可识别性结果，表明在比之前研究的多视角设置更一般的情况下，恢复共享因素是可能的。具体而言，我们区分了具有单一生成机制的多视角设置 (例如，相同类型的多个相机) 和由不同机制特征化的多模态设置 (例如，相机和麦克风)。我们的工作通过重新定义生成过程，考虑具有模态特定潜在变量的不同机制，从而推广了之前的可识别性结果。我们证明了对比学习可以阻塞识别模态之间共享的潜在因素，即使这些因素之间存在非平凡的依赖关系。我们通过数值仿真实证验证了我们的可识别性结果，并在复杂的图像/文本对多模态数据集上证实了我们的发现。放眼全局，我们的工作为多模态表示学习提供了理论基础，并解释了多模态对比学习在实践中可以有效的设置。

1 INTRODUCTION

1 引言

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).

多模态表示学习是一个新兴领域，其发展受到近期弱监督学习算法的进展和适当的多模态数据集收集的推动。多模态数据的特征在于来自两个或多个相关数据源的观察同时发生，例如配对的图像和标题 (例如，Salakhutdinov 和 Hinton，2009；Shi 等，2019；Radford 等，2021)，更一般地说，多模态观察由来自不同类型传感器的对齐测量组成 (Baltrušaitis 等，2019)。共现是一种弱监督的形式 (Shu 等，2020；Locatello 等，2020；Chen 和 Batmanghelich，2020)，因为配对观察可以被视为共享但未观察到的真实因素的代理 (即，弱标签)。在适合弱监督数据的表示学习方法中，对比学习 (Gutmann 和 Hyvärinen，2010；Oord 等，2018) 脱颖而出，因为它旨在利用来自不同视角的共现观察。在实践中，对比学习在多视角和多模态学习中取得了令人鼓舞的结果——一个显著的例子是 CLIP(Radford 等，2021) 对文本到图像生成的突破性进展的贡献 (Ramesh 等，2021；2022；Rombach 等，2022；Saharia 等，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.

尽管对比学习在实践中取得了经验上的成功，但其有效性的解释尚不够清晰。近期的研究将其有效性归因于从潜在因果图中恢复共享潜在因素 (Gresele et al., 2019; Zimmermann et al., 2021; von Kügelgen et al., 2021)。从多视角独立成分分析的角度来看，研究表明对比学习能够逆转应用于具有相互独立成分的潜在变量的非线性混合函数 (即非线性生成过程)(Gresele et al., 2019; Zimmermann et al., 2021)。最近，von Kügelgen et al.(2021) 表明，即使潜在成分之间存在非平凡的因果和统计依赖关系，对比学习仍然能够恢复共享因素，直到块状不确定性。这些结果共同表明，对比学习能够仅通过观察对来识别未知数据生成过程的部分，即使是来自具有非平凡依赖关系的高维多视角观察。在我们的研究中，我们探讨了多模态环境中共享潜在因素的可识别性。

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.

我们考虑一个具有特定模态混合函数和特定模态潜变量的生成过程。我们的设计受到多模态数据固有异质性的启发，当观察是由不同类型的传感器生成时，这种异质性自然出现 (Baltrušaitis 等，2019)。例如，代理可以通过不同的感官模态感知其环境，例如摄像头感知光线或麦克风检测声波。为了建模在模态之间共享的信息，我们从多视角设置 (von Kügelgen 等，2021) 中获得灵感，并允许潜变量之间存在非平凡的依赖关系。然而，以前的工作仅考虑相同数据类型的观察，并假设相同的输入在不同视角下会导致相同的输出。在本研究中，我们引入了一个具有不同生成机制的模型，每个机制都可以表现出显著的模态特定变异。这一区别使得多模态设置相比于以往工作考虑的多视角设置更加一般化。

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

*Joint authorship. † Correspondence to: dimant@ethz.ch.
* 联合作者。† 通信地址:dimant@ethz.ch。

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).

简而言之，我们的工作关注于多模态表示学习的可识别性，并专注于对比学习作为一种特定算法，我们为其推导可识别性结果。在第二节中，我们涵盖了与可识别性和对比学习这两个主题相关的背景知识。然后，我们将多模态生成过程形式化为潜变量模型 (第三节)，并证明对比学习可以阻塞识别模态之间共享的潜在因素 (第四节)。我们通过完全控制的数值模拟 (第五节 1) 实证验证可识别性结果，并在复杂的图像/文本对多模态数据集上证实我们的发现 (第五节 2)。最后，我们将相关文献进行背景化 (第六节)，并讨论潜在的局限性和未来工作的机会 (第七节)。

2 PRELIMINARIES

2 前言

2.1 IDENTIFIABILITY

2.1 可识别性

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).

可识别性是独立成分分析 (ICA)、因果发现和逆问题等领域许多问题的核心 (Lehmann 和 Casella，2006)。从 ICA 的角度来看，我们考虑关系 $\mathbf{x} = \mathbf{f}(\mathbf{z})$ ，其中观察值 $\mathbf{x}$ 是通过应用于潜变量 $\mathbf{z}$ 的混合函数 $\mathbf{f}$ 生成的。ICA 的目标是反转混合函数，以便仅从观察中恢复潜变量。在许多情况下，完全可识别性是不可能的，某些模糊性可能是可以接受的。例如，可识别性可能仅适用于一部分成分 (即部分可识别性)。典型的模糊性包括置换和逐元素变换 (即成分级不确定性)，或对潜变量组的可识别性 (即块级不确定性)。在一般情况下，当 $\mathbf{f}$ 是非线性函数时，一个显著的负结果表明，在给定独立同分布观察的情况下，恢复潜变量在根本上是不可能的 (Hyvärinen 和 Pajunen，1999)。然而，最近的一系列开创性工作在额外假设下为困难的非线性情况提供了可识别性结果，例如辅助变量 (Hyvärinen 和 Morioka，2017；Hyvärinen 等，2019；Khemakhem 等，2020) 或多视角 (Gresele 等，2019；Locatello 等，2020；Zimmermann 等，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; Lyu et al., 2022). Generally, this line of work considers the following generative process:

与我们的研究最相关的是与多视角非线性独立成分分析 (Gresele et al., 2019; Lyu and Fu, 2020; Locatello et al., 2020; von Kügelgen et al., 2021; Lyu et al., 2022) 相关的先前工作。一般而言，这一研究方向考虑以下生成过程:

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

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 considers mutually independent latent components (Song et al., 2014; Gresele et al., 2019; Locatello et al., 2020) or independent groups of shared and view-specific components (Lyu and Fu, 2020; Lyu et al., 2022). Moreover, some of these works (Song et al., 2014; Gresele et al., 2019; Lyu and Fu, 2020; Lyu et al., 2022) consider view-specific [1] mixing functions. Venturing beyond the strict assumption of independent (groups of) components, von Kügelgen et al. (2021) consider additional causal and statistical dependencies between latents 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.

在这里，一个潜变量或其部分组成部分在观察对之间共享 $(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1,\mathbf{x}_2}$ ，而这两个视图 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 是由两个非线性混合函数 $\mathbf{f}_1$ 和 $\mathbf{f}_2$ 生成的。直观上，如果两个视图都包含一个共享信号但在其他方面足够不同，那么第二个视图可以解决由非线性混合引入的模糊性 (Gresele et al., 2019)。之前的研究在对混合函数和潜在组件之间的依赖关系的假设上有所不同。大多数先前的工作考虑了相互独立的潜在组件 (Song et al., 2014; Gresele et al., 2019; Locatello et al., 2020) 或独立的共享和视图特定组件组 (Lyu and Fu, 2020; Lyu et al., 2022)。此外，其中一些工作 (Song et al., 2014; Gresele et al., 2019; Lyu and Fu, 2020; Lyu et al., 2022) 考虑了视图特定的 [1] 混合函数。超越独立 (组) 组件的严格假设，von Kügelgen et al. (2021) 考虑了潜变量之间的额外因果和统计依赖关系，并表明共享组件的子集可以在块状不确定性下被识别。我们的工作考虑了具有因果和统计依赖关系的异质模态。我们证明了共享因子可以在具有模态特定混合函数和模态特定潜变量的新设置中被块识别。

2.2 CONTRASTIVE LEARNING

2.2 对比学习

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:

对比学习 (Gutmann 和 Hyvärinen, 2010; Oord 等, 2018) 是一种自监督表示学习方法，它利用成对观察的弱监督。从高层次来看，该方法学习区分从联合分布中抽样的“正”编码对与从边际乘积中抽样的“负”编码对。流行的 InfoNCE 目标 (Oord 等, 2018) 定义如下:

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

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 $\operatorname{sim}(\cdot,\cdot)$, 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 $I\left(\mathbf{g}_1\left(\mathbf{x}_1\right); \mathbf{g}_2\left(\mathbf{x}_2\right)\right)$ (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 $\operatorname{sim}(a, b) = -(a-b)^2$, it asymptotically behaves like the objective

其中 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 是第一视图和第二视图的编码器，分别为 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 。当 $\mathbf{x}_2$ 是 $\mathbf{x}_1$ 的增强版本，或当从同一分布中抽样的两个增强版本分别用于转换 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 时 (例如，Chen 等, 2020)，通常会使用单个编码器 $\mathbf{g}_1 = \mathbf{g}_2$ 。超参数集包括温度 τ 、相似性度量 $\operatorname{sim}(\cdot,\cdot)$ 和一个整数 K ，该整数控制用于对比的负编码对的数量 (K - 1)。该目标具有信息论解释，作为互信息 $I\left(\mathbf{g}_1\left(\mathbf{x}_1\right); \mathbf{g}_2\left(\mathbf{x}_2\right)\right)$ 的变分下界 (Oord 等, 2018; Poole 等, 2019)，也可以解释为正编码对 (分子) 与额外熵正则化 (分母) 的对齐，其中正则化器抑制了一个退化解，在该解中两个编码器都映射到一个常数 (Wang 和 Isola, 2020)。形式上，当用 $\tau = 1$ 和 $\operatorname{sim}(a, b) = -(a-b)^2$ 实例化 $\mathcal{L}_{\text{InfoNCE}}$ 目标时，它渐近地表现得像该目标。

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

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

对于单个编码器 $\mathbf{g}$，当 $K \to \infty$ (Wang 和 Isola, 2020； von Kügelgen 等, 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., 2022; Radford et al., 2021), which can be obtained by computing the mean of the loss in both directions:

在具有两种异构模态的设置中，自然采用独立的编码器 $\mathbf{g}_1 \neq \mathbf{g}_2$，这可以表示不同的架构。此外，通常使用目标的对称版本 (例如，见 Zhang 等, 2022； Radford 等, 2021)，这可以通过计算两个方向上的损失均值来获得:

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

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:

类似于方程 (3)，我们可以通过大量的负样本 ($K \to \infty$) 来近似 $\tau = 1$ 和 $\text{sim}(a,b) = -(a-b)^2$ 的对称目标，如下所示:

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1,\mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\left\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\right\|_2\right] - 1/2\left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right). \tag{5}$$

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.

由于相似性度量是对称的，因此对齐项的近似对于 $\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_1,\mathbf{g}_2)$ 和 $\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}_2,\mathbf{g}_1)$ 是相同的。每个熵项通过各自损失项的分母进行近似，这可以视为非参数熵估计器 (Wang 和 Isola, 2020)。对于我们的实验，我们采用有限样本估计器 $\mathcal{L}_{\text{InfoNCE}}$ 和 $\mathcal{L}_{\text{SymInfoNCE}}$，而在我们的理论分析中，我们使用估计量 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 来推导可识别性结果。

3 GENERATIVE PROCESS

3 生成过程

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).

在接下来的部分中，我们将多模态生成过程表述为潜变量模型 (第 3.1 节)，然后指定我们对模态之间关系的技术假设 (第 3.2 节)。

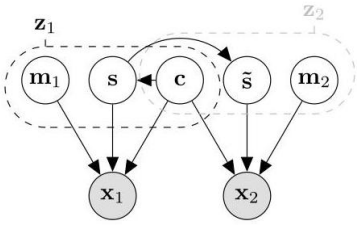


Figure 1: Illustration of the multimodal generative process. Latent variables are denoted by clear nodes and observations by shaded nodes. We partition the latent space into $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 =$

[1] Note that we define modality-specific functions similar to the way Gresele et al. (2019), Lyu and Fu (2020), and Lyu et al. (2022) 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 latents.

[1] 请注意，我们定义了特定于模态的函数，类似于 Gresele 等人 (2019)、Lyu 和 Fu(2020) 以及 Lyu 等人 (2022) 定义特定于视图的函数。为了澄清这种区别，我们通常假设来自不同模态的观察是由不同机制生成的 $\mathbf{f}_1 \neq \mathbf{f}_2$，并且具有特定于模态的潜变量，我们将多视图设置视为一种特殊情况，其中 $\mathbf{f}_1 = \mathbf{f}_2$ 没有特定于视图的潜变量。

$(\widetilde{\mathbf{c}}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$, where $\widetilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 1) and hence we consider only $\mathbf{c}$. Further, $\widetilde{\mathbf{s}}$ 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.

图 1: 多模态生成过程的示意图。潜变量用清晰的节点表示，观察用阴影节点表示。我们将潜空间划分为 $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ 和 $\mathbf{z}_2 = (\widetilde{\mathbf{c}}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$ ，其中 $\widetilde{\mathbf{c}} = \mathbf{c}$ 几乎处处成立 (假设 1)，因此我们仅考虑 $\mathbf{c}$ 。此外，$\widetilde{\mathbf{s}}$ 是 $\mathbf{s}$ 的一个扰动版本 (假设 2)，而 $\mathbf{m}_1, \mathbf{m}_2$ 是特定于模态的变量。观察 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 是通过两个不同的混合函数 $\mathbf{f}_1 \neq \mathbf{f}_2$ 生成的，这些混合函数分别应用于潜变量的子集 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ 。

3.1 LATENT VARIABLE MODEL

3.1 潜变量模型

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:

从高层次来看，我们假设存在一个连续随机变量 $\mathbf{z}$ ，其取值在潜空间 $\mathcal{Z} \subseteq \mathbb{R}^n$ 中，该空间包含生成两种模态观察的所有信息。[2] 此外，我们假设 $\mathbf{z} = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1, \mathbf{m}_2)$ 可以唯一地划分为四个不相交的部分:

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

(i) 一个不变部分 $\mathbf{c}$ ，它在所有模态中始终共享，我们称之为内容;

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

(ii) 一个可变部分 $\mathbf{s}$ ，它可能在不同模态之间变化，我们称之为风格;

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

(iii) 两个特定于模态的部分 $\mathbf{m}_1$ 和 $\mathbf{m}_2$ ，每个部分都是各自模态所独有的。

Let $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 = (\widetilde{\mathbf{c}}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$, where $\widetilde{\mathbf{c}} = \mathbf{c}$ almost everywhere and $\widetilde{\mathbf{s}}$ 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:

设定 $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ 和 $\mathbf{z}_2 = (\widetilde{\mathbf{c}}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$ ，其中 $\widetilde{\mathbf{c}} = \mathbf{c}$ 几乎处处成立，且 $\widetilde{\mathbf{s}}$ 是由第 3.2 节中指定的扰动生成的。类似于多视图独立成分分析 (方程 1)，我们将模态 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的生成过程定义如下:

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

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. We assume that $p_{\mathbf{z}}$ is a smooth density that factorizes as $p_{\mathbf{z}} = p_{\mathbf{c}} p_{\mathbf{s}|\mathbf{c}} p_{\mathbf{m}_1} p_{\mathbf{m}_2}$ in the causal setting, and as the product of all involved marginals when there is no causal dependence from $\mathbf{c}$ to $\mathbf{s}$.

其中 $\mathbf{f}_1 : \mathcal{Z}_1 \to \mathcal{X}_1$ 和 $\mathbf{f}_2 : \mathcal{Z}_2 \to \mathcal{X}_2$ 是两个光滑且可逆的混合函数，具有光滑的逆 (即微分同胚)，它们生成观察值 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ ，分别取值于 $\mathcal{X}_1 \subseteq \mathbb{R}^{d_1}$ 和 $\mathcal{X}_2 \subseteq \mathbb{R}^{d_2}$ 。一般来说，我们假设来自不同模态的观察是由不同机制 $\mathbf{f}_1 \neq \mathbf{f}_2$ 生成的，这些机制以模态特定的潜变量作为输入。至于多视图设置 (von Kügelgen 等，2021)，所考虑的生成过程超越了经典独立成分分析的设置，允许变量块内的统计依赖 (例如，$\mathbf{c}$ 的维度之间) 以及从内容到风格的因果依赖，如图 1 所示。我们假设 $p_{\mathbf{z}}$ 是一个光滑的密度，在因果设置中分解为 $p_{\mathbf{z}} = p_{\mathbf{c}} p_{\mathbf{s}|\mathbf{c}} p_{\mathbf{m}_1} p_{\mathbf{m}_2}$ ，而在没有从 $\mathbf{c}$ 到 $\mathbf{s}$ 的因果依赖时，则作为所有相关边际的乘积。

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 nondeterministic 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).

概述的生成过程相当一般，适用于各种实际设置。内容不变性描述了一种共享现象，该现象并未直接观察到，但在两种模态的观察中表现出来。风格变化描述了在模态间不具鲁棒性的共享影响，例如，数据增强等不可逆变换，或未观察到的混杂因素的非确定性效应。模态特定因素可以视为描述每种模态固有异质性的变量 (例如，背景噪声)。

3.2 RELATION BETWEEN MODALITIES

3.2 模态之间的关系

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., $\widetilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 1), and that $\widetilde{\mathbf{s}}$ 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.

接下来，我们通过定义条件分布 $p_{\mathbf{z}_2|\mathbf{z}_1}$ 来指定我们对模态之间关系的假设，该分布描述了潜在变量 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ 之间的关系，从中通过方程 (6) 生成观测值 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 。与之前在多视角设置中的工作 (von Kügelgen et al., 2021) 类似，我们假设内容是不变的，即 $\widetilde{\mathbf{c}} = \mathbf{c}$ 几乎处处成立 (假设 1)，并且 $\widetilde{\mathbf{s}}$ 是 $\mathbf{s}$ 的扰动版本 (假设 2)。为了陈述我们在多模态设置下的假设，我们还需要考虑特定于模态的潜在变量。

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

假设 1(内容不变性)。条件密度 $p_{\mathbf{z}_2|\mathbf{z}_1}$ 在 $\mathcal{Z}_2 \times \mathcal{Z}_1$ 上的形式为

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

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

对于某个在 $\mathcal{S} \times \mathcal{S}$ 上的连续密度 $p_{\widetilde{\mathbf{s}}|\mathbf{s}}$ ，其中 $\delta\left(\cdot\right)$ 是狄拉克　函数，即 $\widetilde{\mathbf{c}} = \mathbf{c}$ 几乎处处成立。

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

为了完全指定 $p_{\mathbf{z}_2|\mathbf{z}_1}$ ，还需要定义样式变化，这些变化由条件分布 $p_{\widetilde{\mathbf{s}}|\mathbf{s}}$ 描述。对 $\mathbf{s}$ 和 $\widetilde{\mathbf{s}}$ 之间建模这种随机关系有几种理由 (Zimmermann et al., 2021； von Kügelgen et al., 2021)；可以认为 $\widetilde{\mathbf{s}}$ 是 $\mathbf{s}$ 的噪声版本，或者认为 $\widetilde{\mathbf{s}}$ 是一种增强的结果，该增强对 $\mathbf{s}$[3] 施加了软干预。

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

假设 2(风格变化)。设 $\mathcal{A}$ 为风格变量 $\{1, \ldots, n_s\}$ 的幂集，设 p_A 为 $\mathcal{A}$ 上的分布。那么，风格条件 $p_{\widetilde{\mathbf{s}}|\mathbf{s}}$ 是通过对集合 A 进行条件化获得的:

$$p_{\widetilde{\mathbf{s}}|\mathbf{s}}\left(\widetilde{\mathbf{s}} \mid \mathbf{s}\right) = \sum_{A \in \mathcal{A}} p_A\left(A\right) \left(\delta\left(\widetilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}\right) p_{\widetilde{\mathbf{s}}_A|\mathbf{s}_A}\left(\widetilde{\mathbf{s}}_A \mid \mathbf{s}_A\right)\right) \tag{8}$$

where $p_{\widetilde{\mathbf{s}}_A|\mathbf{s}_A}$ 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 $A^c = \{1, \ldots, n_s\} \setminus A$ denotes the complement of A .

其中 $p_{\widetilde{\mathbf{s}}_A|\mathbf{s}_A}$ 是在 $\mathcal{S}_A \times \mathcal{S}_A, \mathcal{S}_A \subseteq \mathcal{S}$ 上的连续密度，表示由 A 指定的变化风格变量的子空间，而 $A^c = \{1, \ldots, n_s\} \setminus A$ 表示 A 的补集。

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

此外，对于任何风格变量 $l \in \{1, \ldots, n_s\}$ ，存在一个集合 $A \subseteq \{1, \ldots, n_s\}$ ，使得 $l \in A$ ，即:

(i) $p_A\left(A\right) > 0$,

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

(ii) $p_{\widetilde{\mathbf{s}}_A|\mathbf{s}_A}$ 对 $\mathbf{s}_A$ 和 $\widetilde{\mathbf{s}}_A$ 都是光滑的，并且

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

(iii) 对于任何 $\mathbf{s}_A, p_{\widetilde{\mathbf{s}}_A|\mathbf{s}_A}\left(\cdot \mid \mathbf{s}_A\right) > 0$ ，在某个包含 $\mathbf{s}_A$ 的开非空子集内。

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 $A \subseteq \{1, \ldots, n_s\}$, which are jointly perturbed to obtain $\widetilde{\mathbf{s}}$. Condition (i) ensures that every style dimension has a positive probability to be perturbed, [4] while (ii) and (iii) are technical smoothness conditions that will be used for the proof of Theorem 1.

直观上，为了生成一对观察值 $(\mathbf{x}_1, \mathbf{x}_2)$ ，我们独立地为每个风格维度翻转一个偏置硬币，以选择一组风格特征 $A \subseteq \{1, \ldots, n_s\}$ ，这些特征共同扰动以获得 $\widetilde{\mathbf{s}}$ 。条件 (i) 确保每个风格维度都有正概率被扰动，[4] ，而 (ii) 和 (iii) 是将用于定理 1 证明的技术光滑性条件。

---

[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.

换句话说，我们假设所有观测值位于一个连续流形上，该流形的维度可能远小于相应模态的观测空间。

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.

总结来说，在本节中，我们将多模态生成过程形式化为潜变量模型 (第 3.1 节)，并通过条件分布 $p_{\mathbf{z}_1|\mathbf{z}_2}$ 指定了我们对模态之间关系的假设 (第 3.2 节)。接下来，我们过渡到表示学习的话题，并展示对于指定的生成过程，多模态对比学习可以识别内容因子，直到块状不确定性。

4 IDENTIFIABILITY RESULTS

4 可识别性结果

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.).

首先，我们需要为多模态设置定义块可识别性 (von Kügelgen et al., 2021)，在该设置中，我们考虑特定于模态的混合函数和编码器。在下文中，n_c 表示内容变量的数量，下标 $1:n_c$ 表示内容维度的子集 (从 1 到 n_c 的索引，且不失一般性)。

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 there exists an invertible function $\mathbf{h}_i : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$, s.t. for the inferred content partition $\widehat{\mathbf{c}}_i = \mathbf{g}_i(\mathbf{x}_i)_{1:n_c}$ it holds that $\widehat{\mathbf{c}}_i = \mathbf{h}_i(\mathbf{c})$.

定义 1(块可识别性)。如果存在一个可逆函数 $\mathbf{h}_i : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ ，使得对于推断的内容分区 $\widehat{\mathbf{c}}_i = \mathbf{g}_i(\mathbf{x}_i)_{1:n_c}$ ，有 $\widehat{\mathbf{c}}_i = \mathbf{h}_i(\mathbf{c})$ ，则真实内容分区 $\mathbf{c} = \mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c}$ 被函数 $\mathbf{g}_i : \mathcal{X}_i \to \mathcal{Z}_i$ 块识别，且 $i \in \{1, 2\}$ 。

It is important to note that block-identifiability does not require the identification of individual factors, which is the goal in 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.

重要的是要注意，块可识别性并不要求识别单个因素，这在多视角非线性独立成分分析 (Gresele et al., 2019; Locatello et al., 2020; Zimmermann et al., 2021; Klindt et al., 2021) 中是目标，也是严格定义解缠结的基础 (Bengio et al., 2013; Higgins et al., 2018; Shu et al., 2020)。相反，我们的目标是将不变因素的组 (即内容分区) 与数据中的其他变异因素隔离开来。

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.

具体而言，我们的目标是证明对比学习可以块识别第 3 节中描述的多模态设置的内容变量。我们在定理 1 中对此进行了形式化，从而通过允许不同的生成机制 $\mathbf{f}_1 \neq \mathbf{f}_2$ 和额外的特定于模态的潜变量来放宽之前工作的假设。

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	x	x	1.00 ± 0.00	0.00 ± 0.00
0.75	X	x	0.99 ± 0.01	0.00 ± 0.00
0.75	✓	X	0.99 ± 0.00	0.52 ± 0.09
0.75	x	✓	1.00 ± 0.00	0.79 ± 0.04
0.75	✓	√	0.99 ± 0.01	0.81 ± 0.04

[3] Note that the asymmetry between $\mathbf{z}_1$ and $\mathbf{z}_2$ (or between $\mathbf{s}$ and $\widetilde{\mathbf{s}}$) 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 perturbations of $\mathbf{z}$, as described in Appendix A.2.

[3] 请注意，$\mathbf{z}_1$ 和 $\mathbf{z}_2$ 之间 (或 $\mathbf{s}$ 和 $\widetilde{\mathbf{s}}$ 之间) 的不对称性并不是严格要求的。我们选择将 $\mathbf{z}_2$ 写作 $\mathbf{z}_1$ 的扰动，以简化符号并与之前的工作保持一致。相反，我们可以通过 $\mathbf{z}$ 的扰动来建模 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ ，如附录 A.2 中所述。

[4] If a style variable would be perturbed with zero probability, it would be a content variable.

[4] 如果一个风格变量的扰动概率为零，则它将是一个内容变量。

生成过程			R^2 (非线性)	
p(chg.)	Stat.	Cau.	内容 c	风格 s
1.0	x	x	1.00 ± 0.00	0.00 ± 0.00
0.75	X	x	0.99 ± 0.01	0.00 ± 0.00
0.75	✓	X	0.99 ± 0.00	0.52 ± 0.09
0.75	x	✓	1.00 ± 0.00	0.79 ± 0.04
0.75	✓	$\surd$	0.99 ± 0.01	0.81 ± 0.04

(a) Original setting
(a) 原始设置

Generative process			R^2 (nonlinear)		
p(chg.)	Stat.	Cau.	Content c	Style s	Modality m_i
1.0	x	x	0.99 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	x	x	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	✓	x	0.95 ± 0.01	0.56 ± 0.23	0.00 ± 0.00
0.75	x	✓	0.98 ± 0.00	0.87 ± 0.04	0.00 ± 0.00
0.75	✓	✓	0.95 ± 0.03	0.89 ± 0.07	0.00 ± 0.00

生成过程			R^2 (非线性)		
p(chg.)	统计	因果	内容 c	风格 s	模态 m_i
1.0	x	x	0.99 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	x	x	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	✓	x	0.95 ± 0.01	0.56 ± 0.23	0.00 ± 0.00
0.75	x	✓	0.98 ± 0.00	0.87 ± 0.04	0.00 ± 0.00
0.75	✓	✓	0.95 ± 0.03	0.89 ± 0.07	0.00 ± 0.00

(b) Multimodal setting
(b) 多模态设置

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.

表 1: 数值模拟的结果。我们将原始设置 ($\mathbf{f}_1 = \mathbf{f}_2$ ，左表) 与多模态设置 (($\mathbf{f}_1 \neq \mathbf{f}_2$ ，右表) 进行比较。只有多模态设置包含特定于模态的潜在变量。每一行展示了不同设置的结果，具有不同的风格变化概率 p(chg.) 和可能的统计 (Stat.) 和/或因果 (Caus.) 依赖关系。每个值表示 R^2 的决定系数 (在 3 个种子上平均) 用于预测相应的真实因素 ($\mathbf{c}, \mathbf{s}$, or $\mathbf{m}_i$ 的非线性回归模型。

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 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 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.

定理 1. 假设数据生成过程如第 3.1 节所述，即数据对 $(\mathbf{x}_1, \mathbf{x}_2)$ 是从方程 (6) 生成的，且 $p_{\mathbf{z}_1} = p_{\mathbf{z} \setminus \{\mathbf{m}_2\}}$ 和 $p_{\mathbf{z}_2|\mathbf{z}_1}$ 如假设 1 和 2 中定义。此外，假设 $p_{\mathbf{z}}$ 是 $\mathcal{Z}$ 上的平滑和连续密度，几乎处处为 $p_{\mathbf{z}}(\mathbf{z}) > 0$ 。设 $\mathbf{g}_1 : \mathcal{X}_1 \to (0,1)^{n_c}$ 和 $\mathbf{g}_2 : \mathcal{X}_2 \to (0,1)^{n_c}$ 为平滑函数，最小化如方程 (5) 中定义的 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 。那么，$\mathbf{g}_1$ 和 $\mathbf{g}_2$ 在定义 1 的意义上区分真实内容变量。

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 5). 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.

定理 1 的证明见附录 A.1。直观上，该结果表明对比学习可以识别内容变量，直到块状不确定性。与之前的工作类似，该结果基于对比损失的渐近形式的优化 (方程 5)。此外，定理 1 假设内容变量的数量是已知的，或者可以估计 (例如，使用肘部法则等启发式方法)。我们在实验中通过维度消融来解决选择编码大小的问题。在第 7 节中，我们将回到实验结果背景下对假设的讨论。

5 EXPERIMENTS

5 实验

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). The code is provided in our github repository. [5]

我们实验的目标是测试对比学习是否能够在多模态设置中进行块识别，如定理 1 所描述的那样。首先，我们在完全受控的环境中通过数值模拟验证可识别性 (第 5.1 节)。其次，我们在一个复杂的图像/文本对多模态数据集上证实我们的发现 (第 5.2 节)。代码已在我们的 GitHub 仓库中提供。[5]

5.1 NUMERICAL SIMULATION

5.1 数值模拟

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 $\mathbf{c} \sim \mathcal{N}\left(0, \sum_{\mathbf{c}}\right), \mathbf{m}_i \sim \mathcal{N}\left(0, \sum_{\mathbf{m}_i}\right)$, and $\mathbf{s} \sim \mathcal{N}\left(\mathbf{a} + B\mathbf{c}, \sum_{\mathbf{s}}\right)$. 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 $\sum_{\mathbf{c}}$). To induce a causal dependence from content to style, we set $a_i, B_{ij} \sim \mathcal{N}(0,1)$; otherwise, we set $a_i, B_{ij} = 0$. For style changes, Gaussian noise is added with probability π independently for each style dimension: $\widetilde{\mathbf{s}}_i = \mathbf{s}_i + \epsilon$, where $\epsilon \sim \mathcal{N}\left(0, \sum_{\epsilon}\right)$ with probability π . We generate the observations $\mathbf{x}_1 = \mathbf{f}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{x}_2 = \mathbf{f}_2(\mathbf{c}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$ 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 activations. We train the encoders for 300,000 iterations using the symmetrized InfoNCE objective (Equation 4) and the hyperparameters listed in Appendix B.1. 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.

我们扩展了 von Kügelgen 等人 (2021) 的数值模拟，并使用特定模态的混合函数 ($\mathbf{f}_1 \neq \mathbf{f}_2$) 和特定模态的潜变量实现了多模态设置。数值模拟使我们能够在对生成过程的完全控制下测量可识别性。数据生成与第 3 节中描述的生成过程一致。我们采样 $\mathbf{c} \sim \mathcal{N}\left(0, \sum_{\mathbf{c}}\right), \mathbf{m}_i \sim \mathcal{N}\left(0, \sum_{\mathbf{m}_i}\right)$ 和 $\mathbf{s} \sim \mathcal{N}\left(\mathbf{a} + B\mathbf{c}, \sum_{\mathbf{s}}\right)$ 。块内的统计依赖性 (例如，$\mathbf{c}$ 的组件之间) 是由相应协方差矩阵中非零的非对角元素引起的 (例如，在 $\sum_{\mathbf{c}}$ 中)。为了引入从内容到风格的因果依赖，我们设置 $a_i, B_{ij} \sim \mathcal{N}(0,1)$ ；否则，我们设置 $a_i, B_{ij} = 0$ 。对于风格变化，以概率 π 独立地为每个风格维度添加高斯噪声: $\widetilde{\mathbf{s}}_i = \mathbf{s}_i + \epsilon$ ，其中 $\epsilon \sim \mathcal{N}\left(0, \sum_{\epsilon}\right)$ 以概率 π 。我们使用两个不同的非线性混合函数生成观察值 $\mathbf{x}_1 = \mathbf{f}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ 和 $\mathbf{x}_2 = \mathbf{f}_2(\mathbf{c}, \widetilde{\mathbf{s}}, \mathbf{m}_2)$ ，即对于每个 $i \in \{1,2\}, \mathbf{f}_i : \mathbb{R}^d \to \mathbb{R}^d$ 是一个单独的、可逆的 3 层 MLP，具有 LeakyReLU 激活函数。我们使用对称化的 InfoNCE 目标 (方程 4) 和附录 B.1 中列出的超参数训练编码器 300,000 次迭代。我们通过使用核岭回归从学习到的表示中预测真实因素来评估块可识别性，并报告在保留数据上的 R^2 决定系数。

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.

结果我们比较了原始设置 ($\mathbf{f}_1 = \mathbf{f}_2$ ，表 1a) 与多模态设置 ($\mathbf{f}_1 \neq \mathbf{f}_2$ ，表 1b)，发现内容在这两种设置中都可以进行块识别，因为内容的预测得分接近于 1，而风格和模态特定信息的预测得分则接近于随机。与之前的研究一致，我们观察到当存在统计和/或因果依赖时，某些风格信息可以被预测；这是可以预期

[5] https://github.com/imantdaunhawer/multimodal-contrastive-learning.

[5] https://github.com/imantdaunhawer/multimodal-contrastive-learning.

的，因为统计依赖降低了内容的有效维度，而因果依赖 $\mathbf{c} \to \mathbf{s}$ 使得风格可以部分地从编码的内容信息中预测。总体而言，数值模拟的结果与我们从定理 1 得出的理论结果一致，表明对比学习可以在多模态设置中进行内容的块识别。

5.2 IMAGE/TEXT PAIRS

5.2 图像/文本对

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).

接下来，我们测试在更现实的设置中，图像/文本对的块识别性是否成立——这两种复杂模态具有不同的生成机制。我们扩展了 Causal3DIdent 数据集 (von Kügelgen 等，2021；Zimmermann 等，2021)，该数据集允许我们测量和控制用于生成复杂观察的真实潜在因素。我们使用 Blender(Blender 在线社区，2018) 渲染高维图像，描绘一个被不同颜色聚光灯照亮并位于有色背景前的彩色物体的场景。该场景由 11 个潜在因素定义: 物体的形状 (7 个类别)、物体的位置 (x, y, z 坐标)、物体的方向 (α, β, γ angles) 、聚光灯的位置 (θ angle) ，以及物体、背景和聚光灯的颜色 (每个都有一个数值)。

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 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 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. Further details about the dataset are provided in Appendix B.1.

Multimodal3DIdent 我们将 Causal3DIdent 数据集扩展到多模态设置，如下所示。我们通过调整来自 CLEVR 数据集 (Johnson et al., 2017) 的文本渲染，从潜在因素生成文本描述。每对图像/文本共享关于物体形状 (如牛、茶壶等) 及其在场景中的位置 (例如，右下角) 的信息。对于每个位置因素，我们使用三个明显可区分的值 (上/中/下；左/中/右)，这些值在文本中比坐标更自然地描述。虽然形状和位置总是共享 (即内容) 在配对的图像和文本之间，但物体的颜色是由位置因果影响的，并且是随机共享的 (即风格)。对于物体颜色，我们使用连续的色调值，而对于文本，我们将 RGB 值与给定调色板中最近的值匹配 (即一系列命名颜色，如棕色、米色、橄榄色等)。调色板是从三种调色板的集合中随机抽样，以确保图像中描绘的物体颜色不会唯一决定文本中描述的颜色。作为图像的模态特定因素，我们有物体旋转、聚光灯位置和背景颜色，而对于文本描述，我们遵循 Johnson et al. (2017) 的方法，使用 5 种不同类型的短语来引入模态特定的变化。图像/文本对的示例如图 2 所示。关于数据集的更多细节在附录 B.1 中提供。

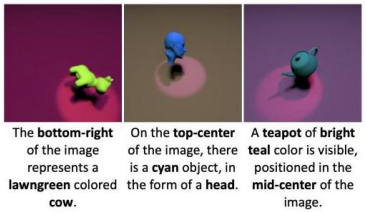


Figure 2: Examples of image/text pairs.

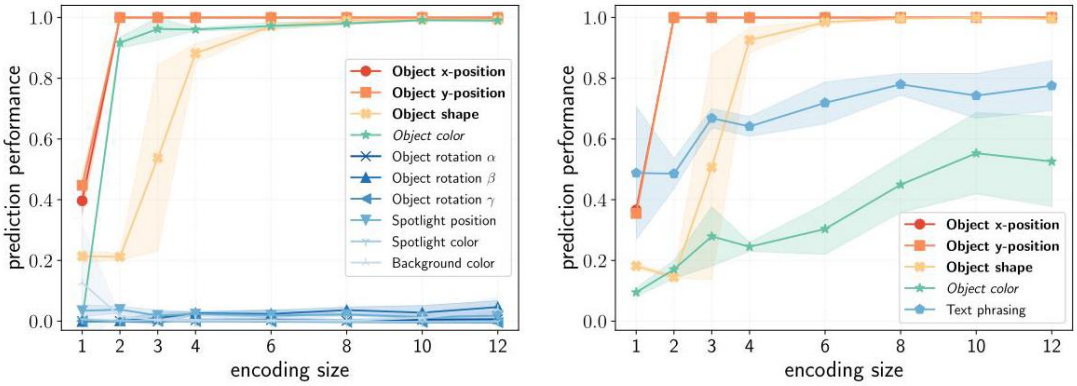
图 2: 图像/文本对的示例。

We train the encoders for 100,000 iterations using the symmetrized InfoNCE objective (Equation 4) and the hyperparameters listed in Appendix B.1. For the image encoder we use a ResNet-18 architecture (He et al., 2016) and for the text we use a convolutional network. 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.

我们使用对称化的 InfoNCE 目标 (方程 4) 和附录 B.1 中列出的超参数训练编码器 100,000 次迭代。对于图像编码器，我们使用 ResNet-18 架构 (He 等，2016)，而对于文本，我们使用卷积网络。至于数值模拟，我们通过从学习到的表示中预测真实因素来评估块可识别性。对于连续因素，我们使用核岭回归并报告 R^2 分数，而对于离散因素，我们报告具有单个隐藏层的多层感知机 (MLP) 的分类准确率。

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 3-4 dimensions). 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

结果图 3 展示了在 Multimodal3DIdent 上的结果，进行了维度消融实验，我们改变了模型编码的大小。内容因素 (物体位置和形状) 始终编码良好，除非编码大小过小 (即小于 3-4 维)。当容量足够时，风格信息 (物体颜色) 也被编码，这部分是因为内容与风格之间存在因果依赖，部分是由于过剩的容量，正如之前的研究所观察到的那样。图像特定信息 (物体旋转、聚光灯位置、背景颜色) 在很大程度上被丢弃，与编码大小无关。文本特定信息 (措辞) 被适度编码 ($48 - 80\%$ accuracy) ，我们将其归因于措辞是一个违反连续潜变量假设的离散因素。这暗示了可能存在的限制。



(a) Prediction of image factors

(b) Prediction of text factors

Figure 3: Results 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.

图 3:Multimodal3DIdent 的结果作为模型编码大小的函数。我们评估真实图像因子 (左子图) 和文本因子 (右子图) 的非线性预测，以量化学习到的表示如何编码各自的因子。内容因子用粗体表示，风格因子用斜体表示。在 x 轴上，我们改变编码大小，即模型的输出维度。我们分别以连续因子的 R^2 决定系数和离散因子的分类准确率来衡量预测性能。每个点表示三个种子的平均值，带状图显示一个标准差。

presence 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.

离散潜在因子的存在，我们在附录 B.2 中进一步探讨，并在第 7 节中讨论。总体而言，我们的结果表明，对比学习可以在复杂的多模态设置中通过图像/文本对阻止识别内容因子。

6 RELATED WORK

6 相关工作

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; Locatello et al., 2020) that recover the latent variable up to a component-wise indeterminacy in a setting with mutually independent latent components, or up to block-wise inderterminacies in the case of independent groups of shared and view-specific components (Lyu and Fu, 2020; Lyu et al., 2022). Beyond the assumption of independent (groups of) 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 heterogeneous 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.

多视角非线性独立成分分析多视角非线性独立成分分析的目标是识别不同视角之间共享的潜在因素，如第 2.1 节所述。有一系列研究 (Gresele et al., 2019; Locatello et al., 2020) 在具有相互独立的潜在成分的情况下，恢复潜在变量，直到成分级别的不确定性，或者在共享和视角特定成分的独立组的情况下，恢复到块级别的不确定性 (Lyu and Fu, 2020; Lyu et al., 2022)。超越独立 (组) 成分的假设，还有一系列研究 (von Kügelgen et al., 2021; Kong et al., 2022) 将潜在空间划分为不变成分块和变化成分块，并表明即使在潜在成分之间存在非平凡的依赖关系时，不变成分也可以在块级别的不确定性下被识别。我们的工作在这个方向上有所进展，考虑了具有潜在变量之间非平凡统计和因果依赖关系的异质模态。我们证明了共享因素可以在具有模态特定混合函数和模态特定潜在变量的新颖设置中被块识别。

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; Bachman et al., 2019; Federici et al., 2020; Tsai et al., 2021; 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.

多模态表示学习多模态表示学习旨在将来自异构来源的信息整合为联合表示 (Baltrušaitis et al., 2019; Guo et al., 2019)。有许多方法旨在直接或间接地学习多模态数据的表示。在间接学习表示的方法中，有多模态自编码器 (Ngiam et al., 2011; Geng et al., 2022; Bachmann et al., 2022; Aghajanyan et al., 2022) 和多种多模态生成模型 (例如，Suzuki et al., 2016; Wu and Goodman, 2018; Shi et al., 2019; Huang et al., 2018; Tsai et al., 2019; Ramesh et al., 2021)，这些模型通过不同形式的重构和/或掩蔽预测误差的反向传播来学习表示。更直接的方法是无解码器的方法，这些方法最大化不同模态编码之间的相似性。这类方法包括非线性典型相关分析 (Akaho, 2001; Bach and Jordan, 2002; Andrew et al., 2013; Wang et al., 2016; Tang et al., 2017; Karami and Schuurmans, 2021)，以及多视角和多模态对比学习 (Tian et al., 2019; Bachman et al., 2019; Federici et al., 2020; Tsai et al., 2021; Radford et al., 2021; Poklukar et al., 2022)。虽然所有这些方法都旨在整合跨模态的信息，但它们并没有回答可识别性这一基本问题，而我们的工作正是旨在解决这一问题。

7 DISCUSSION

7 讨论

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 block-identifies 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. In Appendix B.2, we include further experiments that test violations with discrete factors and dimensionality ablations that examine the robustness and sample complexity. 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 suggest 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 learn a representation that recovers 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.

含义和范围我们已经表明，对比学习可以在多模态环境中块识别共享因素。数值模拟 (第 5.1 节) 验证了我们的主要理论结果 (定理 1)，显示当编码的大小与内容因素的数量匹配时，对比学习能够块识别内容信息 (定义 1)。在一个复杂的图像/文本对数据集上的实验证实，对比学习能够在更现实的环境中隔离内容，甚至在某些违反定理 1 基本假设的情况下也是如此。在附录 B.2 中，我们包括了进一步的实验，测试离散因素的违反和维度消融，考察其鲁棒性和样本复杂性。更一般地说，我们观察到对比学习在所有设置中都能很好地编码不变信息 (即内容)。当容量足够时，随机共享信息 (即风格) 被适度编码，但不会影响不变信息的预测。在实践中，我们的结果表明，在没有容量限制的情况下，对比学习可以编码任何共享因素，无论该因素是否在模态之间真正不变，或其对观察结果的影响是否受到噪声或其他因素的混淆。这与信息论视角一致 (Oord et al., 2018; Poole et al., 2019)，即对比学习最大化表示之间的互信息——一种量化共享信息的互依赖度量。我们的结果表明，编码的大小可以减少，以学习恢复不变信息的表示，这由块可识别性的概念所捕捉。在实践中，这可以在内容保持的分布转移环境中 (Mitrovic et al., 2021; Federici et al., 2021) 用于表示学习，在这些环境中，与下游任务相关的信息保持不变。

Limitations and outlook First, Theorem 1 suggests 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; Lachapelle 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. We explore the idea of estimating the number of content factors in Appendix B. 2 Figure 7. Third, Theorem 1 assumes that all latent factors are continuous. While this assumption prevails in related work (Hyvärinen and Pajunen, 1999; Hyvärinen 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 Figure 5, we provide numerical simulations that support 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) as well as suitable learning objectives (Tian et al., 2019; Lyu et al., 2022). Summarizing, the described limitations mirror the assumptions on the generative process (Section 3), which may be relaxed in future work.

限制与展望首先，定理 1 表明只有不变因素可以被块识别。然而，在实践中，可能存在观察对，其中

不变性被无意中违反，例如，由于测量误差、遮挡或数据收集中的其他错误。一方面，这种违反可以被视为数据收集的一个简单伪影，并可以通过对生成过程的干预来管理，例如，在强化学习中的行动 (Lippe et al., 2022; Brehmer et al., 2022; Ahuja et al., 2022; Lachapelle et al., 2022)。另一方面，内容不变性的违反模糊了内容因素和风格因素之间的界限，研究在只有随机共享因素的更一般环境中的可识别性将是有趣的。其次，定理 1 假设内容因素的数量是已知的或可以被估计。在实践中，这可能不是一个显著的限制，因为内容因素的数量可以被视为一个单一的超参数 (例如，Locatello et al., 2020)，尽管设计合适的启发式方法是一个有趣的研究方向。我们在附录 B.2 中探讨了估计内容因素数量的想法。第三，定理 1 假设所有潜在因素都是连续的。虽然这一假设在相关工作中占主导地位 (Hyvärinen 和 Pajunen, 1999; Hyvärinen 和 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)，但我们在图 3b 中的结果表明，在存在离散因素的情况下，某些风格或模态特定的信息可以被编码。在附录 B.2 的图 5 中，我们提供了支持这些发现的数值模拟。最后，我们的模型可以扩展到两个以上的模态——这是一个有趣的可识别性结果 (Gresele et al., 2019; Schölkopf et al., 2016) 以及合适的学习目标 (Tian et al., 2019; Lyu et al., 2022)。总之，所描述的限制反映了对生成过程的假设 (第 3 节)，这些假设可能在未来的工作中被放宽。

8 CONCLUSION

8 结论

We addressed the problem of identifiability for multimodal representation learning and showed that contrastive learning can block-identify latent factors shared between heterogeneous 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.

我们解决了多模态表示学习中的可识别性问题，并展示了对比学习可以区分异构模态之间共享的潜在因素。我们将多模态生成过程形式化为一种新颖的潜变量模型，该模型具有特定于模态的生成机制以及潜变量之间非平凡的统计和因果依赖关系。我们证明了对比学习可以识别共享的潜在因素，直到块状不确定性，因此可以将模态之间的不变性与其他可变因素隔离开来。我们的理论结果得到了数值仿真和复杂的图像/文本对多模态数据集的验证。更一般地说，我们相信我们的工作将有助于为多模态表示学习塑造理论基础，并且对所提出的生成过程的进一步放宽提供了丰富的未来研究机会。

ACKNOWLEDGEMENTS

致谢

ID was supported by the SNSF grant #200021-188466. EP was supported by the grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). Experiments were performed on the ETH Zurich Leonhard cluster. Special thanks to Kieran Chin-Cheong for his support in the early stages of the project as well as to Luigi Gresele and Julius von Kügelgen for helpful discussions.

ID 得到了 SNSF 资助 #200021-188466 的支持。EP 得到了 ETH 域 (瑞士联邦理工学院) "个性化健康与相关技术 (PHRT)" 战略重点领域资助 #2021-911 的支持。实验是在 ETH 苏黎世的 Leonhard 集群上进行的。特别感谢 Kieran Chin-Cheong 在项目早期阶段的支持，以及 Luigi Gresele 和 Julius von Kügelgen 的有益讨论。

REPRODUCIBILITY STATEMENT

可重复性声明

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

ablations; 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. Additionally, we publish the code to reproduce the experiments.

对于我们的理论陈述，我们提供详细的推导并说明必要的假设。生成过程在第 3 节中指定，块可识别性的假设在定理 1 中引用。我们通过适当的实验 (维度消融；离散潜在因子) 测试关键假设的违反，并在第 7 节中讨论我们工作的局限性。此外，我们通过数值仿真和复杂的多模态数据经验验证我们的理论结果。为了确保经验的可重复性，每个实验的结果在多个种子上进行了平均，并报告了标准偏差。关于实现细节、超参数设置和评估指标的信息包含在附录 B.1 中。此外，我们发布了重现实验的代码。

REFERENCES

参考文献

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. In Advances in Neural Information Processing Systems.

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.

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. Advances in neural information processing systems.

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. In Advances in Neural Information Processing Systems.

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.

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.

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

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.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer.

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.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. (2022). Disentanglement via mechanism sparsity regularization: a new principle for nonlinear ICA. In Conference on Causal Learning and Reasoning.

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.

Lyu, Q., Fu, X., Wang, W., and Lu, S. (2022). Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In International Conference on Learning Representations.

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 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. In Advances in Neural Information Processing Systems.

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. (2022). Contrastive learning of medical visual representations from paired images and text. In Machine Learning for Healthcare.

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

理论

A.1 Proof of Theorem 1

A.1 定理 1 的证明

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 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 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.

定理 1. 假设数据生成过程如第 3.1 节所述，即数据对 $(\mathbf{x}_1, \mathbf{x}_2)$ 是从方程 (6) 生成的，且 $p_{\mathbf{z}_1} = p_{\mathbf{z} \setminus \{\mathbf{m}_2\}}$ 和 $p_{\mathbf{z}_2|\mathbf{z}_1}$ 如假设 1 和 2 中定义。此外，假设 $p_{\mathbf{z}}$ 是在 $\mathcal{Z}$ 上的光滑和连续密度，且在几乎所有地方都有 $p_{\mathbf{z}}(\mathbf{z}) > 0$ 。设 $\mathbf{g}_1 : \mathcal{X}_1 \to (0,1)^{n_c}$ 和 $\mathbf{g}_2 : \mathcal{X}_2 \to (0,1)^{n_c}$ 是光滑函数，最小化如方程 (5) 中定义的 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 。那么，$\mathbf{g}_1$ 和 $\mathbf{g}_2$ 在定义 1 的意义上块可识别真实内容变量。

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.

证明。为了证明定理 1，我们遵循 von Kügelgen 等人 (2021 年，定理 4.4) 的证明结构，并将证明分为三个步骤。首先，我们展示存在一对光滑函数 $\mathbf{g}_1^*, \mathbf{g}_2^*$ ，它们达到 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 的全局最小值 (方程 5)。进一步地，在方程 (13-15) 中，我们推导出必须几乎肯定对任何一对光滑函数 $\mathbf{g}_1, \mathbf{g}_2$ 成立的不变性条件，这些函数达到方程 (5) 的全局最小值。在步骤 2 中，我们利用步骤 1 中推导的不变性条件，通过反证法展示任何一对光滑函数 $\mathbf{g}_1, \mathbf{g}_2$ 只能依赖于内容，而不依赖于风格或特定于模态的信息。在第三个也是最后一个步骤中，对于 $\mathbf{h}_1 := \mathbf{g}_1 \circ \mathbf{f}_1$ 和 $\mathbf{h}_2 := \mathbf{g}_2 \circ \mathbf{f}_2$ ，我们展示这两个函数必须是双射，因此 $\mathbf{c}$ 分别由 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 块识别。

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

步骤 1。回顾目标的渐近形式，如方程 (5) 所定义:

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\left\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\right\|_2\right] - 1/2\left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right). \quad (5)$$

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. [6]

当第一个项最小化且第二个项最大化时，达到 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 的全局最小值。第一个项在编码器 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 完全对齐时最小化，即当 $\mathbf{g}_1(\mathbf{x}_1) = \mathbf{g}_2(\mathbf{x}_2)$ 对所有对 $(\mathbf{x}_1, \mathbf{x}_2) \sim p_{\mathbf{x}_1,\mathbf{x}_2}$ 成立时。第二个项在 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 分别映射到 $(0,1)^{n_c}$ 上的均匀分布随机变量时达到最大值。[6]

To show that there exists a pair of functions that minimize $\mathcal{L}_{\text{SymAlignMaxEnt}}$, let $\mathbf{g}_1^* := \mathbf{d}_1 \circ \mathbf{f}_{1,1:n_c}^{-1}$ and let $\mathbf{g}_2^* := \mathbf{d}_2 \circ \mathbf{f}_{2.1:n_c}^{-1}$, where the subscript $1:n_c$ indexes the subset of content dimensions w.l.o.g. and where $\mathbf{d}_1$ and $\mathbf{d}_2$ 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 $\mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c} = \widetilde{\mathbf{c}}$ by definition. Second,, for $i \in \{1,2\}$, let us define $\mathbf{d}_i : \mathcal{C} \mapsto (0,1)^{n_c}$ using the Darmois construction, such that $\mathbf{d}_i$ maps $\mathbf{c}$ and $\widetilde{\mathbf{c}}$ to a uniform random variable respectively. It follows that $\mathbf{g}_1^*, \mathbf{g}_2^*$ are smooth functions, because any function $\mathbf{d}_i$ 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).

为了证明存在一对函数可以最小化 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ ，设 $\mathbf{g}_1^* := \mathbf{d}_1 \circ \mathbf{f}_{1,1:n_c}^{-1}$ 并设 $\mathbf{g}_2^* := \mathbf{d}_2 \circ \mathbf{f}_{2.1:n_c}^{-1}$ ，其中下标 $1:n_c$ 索引内容维度的子集，且 $\mathbf{d}_1$ 和 $\mathbf{d}_2$ 将使用达尔莫伊斯构造定义 (Darmois, 1951; Hyvärinen 和 Pajunen, 1999)。首先，回顾 $\mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{c}$ ，并且根据定义 $\mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c} = \widetilde{\mathbf{c}}$ 。其次，对于 $i \in \{1,2\}$ ，我们定义 $\mathbf{d}_i : \mathcal{C} \mapsto (0,1)^{n_c}$ 使用达尔莫伊斯构造，使得 $\mathbf{d}_i$ 分别将 $\mathbf{c}$ 和 $\widetilde{\mathbf{c}}$ 映射到均匀随机变量。因此，$\mathbf{g}_1^*, \mathbf{g}_2^*$ 是光滑函数，因为通过达尔莫伊斯构造获得的任何函数 $\mathbf{d}_i$ 都是光滑的，并且 $\mathbf{f}_1^{-1}, \mathbf{f}_2^{-1}$ 也是光滑的 (每个都是光滑函数的逆)。

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

接下来，我们证明上述定义的函数对 $\mathbf{g}_1^*, \mathbf{g}_2^*$ 达到了目标 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 的全局最小值。我们有

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1^*, \mathbf{g}_2^*) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\left\|\mathbf{g}_1^*(\mathbf{x}_1) - \mathbf{g}_2^*(\mathbf{x}_2)\right\|_2\right] - 1/2\left(H(\mathbf{g}_1^*(\mathbf{x}_1)) + H(\mathbf{g}_2^*(\mathbf{x}_2))\right) \quad (9)$$

$$= \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\left\|\mathbf{d}_1(\mathbf{c}) - \mathbf{d}_2(\widetilde{\mathbf{c}})\right\|_2\right] - {}^1/_2\left(H(\mathbf{d}_1(\mathbf{c})) + H(\mathbf{d}_2(\widetilde{\mathbf{c}}))\right) \quad (10)$$

$$= 0 \tag{11}$$

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

其中根据假设 $1, \mathbf{c} = \widetilde{\mathbf{c}}$ 几乎肯定，这意味着第一项几乎肯定为零。此外，$\mathbf{d}_i$ 将 $\mathbf{c}, \widetilde{\mathbf{c}}$ 映射到 $(0,1)^{n_c}$ 上均匀分布的随机变量，这也意味着 $\mathbf{d}_1(\mathbf{c})$ 和 $\mathbf{d}_2(\widetilde{\mathbf{c}})$ 的微分熵为零。因此，存在一对函数 $\mathbf{g}_1^*, \mathbf{g}_2^*$ 可以最小化 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 。

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.,

接下来，设 $\mathbf{g}_1 : \mathcal{X}_1 \mapsto (0,1)^{n_c}$ 和 $\mathbf{g}_2 : \mathcal{X}_2 \mapsto (0,1)^{n_c}$ 为任何一对光滑函数，它们达到了方程 (5) 的全局最小值，即，

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{g}_1, \mathbf{g}_2) = \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p_{\mathbf{x}_1,\mathbf{x}_2}}\left[\left\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\right\|_2\right] - 1/2\left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right) = 0. \tag{12}$$

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

设 $\mathbf{h}_1 := \mathbf{g}_1 \circ \mathbf{f}_1$ 和 $\mathbf{h}_2 := \mathbf{g}_2 \circ \mathbf{f}_2$ ，并注意到由于所有相关函数根据定义都是光滑的，因此两者都是光滑函数。由于方程 (12) 是全局最小值，这意味着以下内容

invariance conditions for the individual terms:

个别项的不变性条件:

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

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

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

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 $\widehat{\mathbf{c}}_1 = \mathbf{h}_1(\mathbf{z}_1)$ (resp. $\widehat{\mathbf{c}}_2 = \mathbf{h}_2(\mathbf{z}_2)$) must be uniform on $(0,1)^{n_c}$.

因此，$\mathbf{h}_1(\mathbf{z}_1) = \mathbf{h}_2(\mathbf{z}_2)$ 必须几乎肯定地相对于 $p_{\mathbf{x}_1,\mathbf{x}_2}$ 成立。此外，方程 (14)(分别为方程 (15)) 意味着 $\widehat{\mathbf{c}}_1 = \mathbf{h}_1(\mathbf{z}_1)$ (分别为 $\widehat{\mathbf{c}}_2 = \mathbf{h}_2(\mathbf{z}_2)$) 必须在 $(0,1)^{n_c}$ 上是均匀的。

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.

第 2 步。接下来，我们展示任何一对最小化 $\mathcal{L}_{\text{SymAlignMaxEnt}}$ 的函数仅依赖于内容信息。由于风格与 $\mathbf{m}_1$ 和 $\mathbf{m}_2$ 无关，我们首先展示 $\mathbf{h}_1(\mathbf{z}_1)$ 不依赖于 $\mathbf{m}_1$ ，并且 $\mathbf{h}_2(\mathbf{z}_2)$ 不依赖于 $\mathbf{m}_2$ 。然后，我们基于先前工作的结果展示 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 也不能依赖于风格。

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 $\mathbf{m}_1$ in the sense that the partial derivative of $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ w.r.t. some modality-specific variable $m_{1,l}$ is non-zero for some point $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*) \in \mathcal{Z}_1$. Specifically, it implies that the partial derivative $\partial \mathbf{h}_1(\mathbf{z}_1)_{1:n_c} / \partial m_{1,l}$ is positive in a neighborhood around $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$, which is a non-empty open set, since $\mathbf{h}_1$ is smooth. On the other hand, due to the independence of $\mathbf{z}_2$ and $\mathbf{m}_1$, the fact that $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ cannot not depend on $\mathbf{m}_1$, and that $p(\mathbf{z}) > 0$ almost everywhere, we come to a contradiction. That is, there exists an open set of points with positive measure, namely the neighbourhood around $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$, on which

[6] 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 $[a,b]$.

[6] 请注意，我们通过定义将 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 的范围限制为 $(0,1)^{n_c}$ 仅仅是为了简化符号。一般来说，均匀分布 $\mathcal{U}(a,b)$ 是区间 $[a,b]$ 上的最大熵分布。

首先注意，我们可以排除所有退化解，其中 $\mathbf{g}_1$ 将 $\mathbf{m}_1$ 的一个分量映射到一个常数，因为 $\mathbf{g}_1$ 将不再是可逆的，这样的解将违反方程 (14) 中的不变性。为了证明矛盾，假设在不失一般性的情况下，$\mathbf{h}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)_{1:n_c} := \mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ 依赖于 $\mathbf{m}_1$ 中的某个分量，意味着某个模态特定变量 $m_{1,l}$ 的偏导数 $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ 在某个点 $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*) \in \mathcal{Z}_1$ 处非零。具体来说，这意味着偏导数 $\partial \mathbf{h}_1(\mathbf{z}_1)_{1:n_c} / \partial m_{1,l}$ 在 $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$ 附近是正的，这是一个非空的开集，因为 $\mathbf{h}_1$ 是光滑的。另一方面，由于 $\mathbf{z}_2$ 和 $\mathbf{m}_1$ 的独立性，$\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ 不能依赖于 $\mathbf{m}_1$ 的事实，以及 $p(\mathbf{z}) > 0$ 几乎处处成立，我们得出了矛盾。也就是说，存在一个正测度的开点集，即 $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$ 附近的邻域。

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

almost surely, which contradicts the invariance in Equation (13). The statement does not change, if we add further dependencies of $\mathbf{h}_1$ on components of $\mathbf{m}_1$, or for $\mathbf{h}_2$ on components of $\mathbf{m}_2$, because $\mathbf{m}_1$ and $\mathbf{z}_2$ are independent, and $\mathbf{m}_2$ and $\mathbf{z}_1$ are independent as well. Hence, we show that any encoder that minimizes the objective in Equation (5) cannot depend on modality-specific information.

几乎肯定，这与方程 (13) 中的不变性相矛盾。如果我们增加 $\mathbf{h}_1$ 对 $\mathbf{m}_1$ 组件的进一步依赖，或 $\mathbf{h}_2$ 对 $\mathbf{m}_2$ 组件的依赖，声明也不会改变，因为 $\mathbf{m}_1$ 和 $\mathbf{z}_2$ 是独立的，而 $\mathbf{m}_2$ 和 $\mathbf{z}_1$ 也是独立的。因此，我们证明了任何最小化方程 (5) 中目标的编码器都不能依赖于模态特定信息。

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 $\widetilde{\mathbf{s}}$, 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).

已经确定 $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ 和 $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ 都不能依赖于特定于模态的信息，接下来需要证明风格信息也没有被编码。利用假设 2，我们可以表明，如果 $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ 或 $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ 依赖于 $\mathbf{s}$ 中的某个维度，则方程 (13) 中的严格不等式具有正密度 $\widetilde{\mathbf{s}}$ ，这将再次导致违反方程 (13) 中推导出的不变性，如 von Kügelgen 等人 (2021，定理 4.2 的证明) 所示。

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)

第 3 步。接下来需要证明 $\mathbf{h}_1, \mathbf{h}_2$ 是双射。我们知道 $\mathcal{C}$ 和 $(0,1)^{n_c}$ 是简单连通且有向的 C^1 流形，并且我们在第 1 步中已经建立了 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 是光滑的，因此是可微分的函数。由于 $p_{\mathbf{c}}$ 是一个常规密度，因此相对于推送函数 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 的均匀分布是常规密度。因此，$\mathbf{h}_1$ 和 $\mathbf{h}_2$ 是双射 (Zimmermann 等人，2021，命题 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 $\mathbf{g}_2$.

第 3 步结束了证明。我们已经证明，对于任何一对光滑函数 $\mathbf{g}_1, \mathbf{g}_2$ ，它们达到方程 (5) 的全局最小值，我们有 $\mathbf{c}$ 被 $\mathbf{g}_1$ 和 $\mathbf{g}_2$ 块识别 (定义 1)。

A.2 SYMMETRIC GENERATIVE PROCESS

A.2 对称生成过程

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 $\mathbf{z}$.

在论文的主体部分，我们描述了一种不对称生成机制，其中 $\mathbf{z}_2$ 是 $\mathbf{z}_1$ 的一个扰动版本。在这里，我们将简要概述我们的模型和结果如何适应对称设置，其中 $\mathbf{z}_1$ 和 $\mathbf{z}_2$ 都是作为 $\mathbf{z}$ 的扰动生成的。

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 $\mathbf{z}_1 = (\widetilde{\mathbf{c}}_1, \widetilde{\mathbf{s}}_1, \widetilde{\mathbf{m}}_1)$ and $\mathbf{z}_2 = (\widetilde{\mathbf{c}}_2, \widetilde{\mathbf{s}}_2, \widetilde{\mathbf{m}}_2)$ are generated.

具体而言，我们需要对假设 1 和 2 进行如下小调整。我们从假设 1 中的内容不变性开始，该假设指定了 $\mathbf{z}_1 = (\widetilde{\mathbf{c}}_1, \widetilde{\mathbf{s}}_1, \widetilde{\mathbf{m}}_1)$ 和 $\mathbf{z}_2 = (\widetilde{\mathbf{c}}_2, \widetilde{\mathbf{s}}_2, \widetilde{\mathbf{m}}_2)$ 的生成方式。

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

设 $i \in \{1, 2\}$ 。条件密度 $p_{\mathbf{z}_i|\mathbf{z}}$ 在 $\mathcal{Z}_i \times \mathcal{Z}$ 上的形式为

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

where $\delta(\cdot)$ is the Dirac delta function, i.e., $\widetilde{\mathbf{c}}_i = \mathbf{c}$ almost everywhere, as well as $\widetilde{\mathbf{m}}_i = \mathbf{m}_i$ almost everywhere. Note that since $\widetilde{\mathbf{c}}_1 = \mathbf{c}$ a.e. and $\mathbf{c} = \widetilde{\mathbf{c}}_2$ a.e., it follows that $\widetilde{\mathbf{c}}_1 = \widetilde{\mathbf{c}}_2$ almost everywhere, which is a property that is needed in Step 1 in the proof of Theorem 1. In addition, it still holds that $\widetilde{\mathbf{m}}_i \perp \mathbf{z}_j$, 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.

其中 $\delta(\cdot)$ 是狄拉克 函数，即 $\widetilde{\mathbf{c}}_i = \mathbf{c}$ 几乎处处成立，以及 $\widetilde{\mathbf{m}}_i = \mathbf{m}_i$ 几乎处处成立。注意，由于 $\widetilde{\mathbf{c}}_1 = \mathbf{c}$ 几乎处处成立和 $\mathbf{c} = \widetilde{\mathbf{c}}_2$ 几乎处处成立，因此 $\widetilde{\mathbf{c}}_1 = \widetilde{\mathbf{c}}_2$ 几乎处处成立，这是在定理 1 的证明步骤 1 中所需的性质。此外，对于 $i, j \in \{1, 2\}$ 和 $i \neq j$ ，仍然成立 $\widetilde{\mathbf{m}}_i \perp \mathbf{z}_j$ ，这在证明的步骤 2 中是必要的，以表明特定模态的信息未被编码。

Lastly, we need to revisit Assumption 2, for which both $\widetilde{\mathbf{s}}_1$ and $\widetilde{\mathbf{s}}_2$ would be generated through perturbations of $\mathbf{s}$ via the conditional distribution $p_{\widetilde{\mathbf{s}}_i|\mathbf{s}}$ 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 $\widetilde{\mathbf{s}}_1$ and $\widetilde{\mathbf{s}}_2$ such that they are not perturbed in an identical manner w.r.t. s. This can be ensured by, e.g., constraining p_A appropriately to exclude the degenerate case where dimensions in $\widetilde{\mathbf{s}}_1$ and $\widetilde{\mathbf{s}}_2$ are perfectly aligned - a case that needs to be excluded for Step 2 of the proof of Theorem 1.

最后，我们需要重新审视假设 2，其中 $\widetilde{\mathbf{s}}_1$ 和 $\widetilde{\mathbf{s}}_2$ 将通过 $\mathbf{s}$ 的扰动生成，依据条件分布 $p_{\widetilde{\mathbf{s}}_i|\mathbf{s}}$ 在 $\mathcal{S} \times \mathcal{S}$ 上，如假设 2 所述，针对每个 i 单独进行。作为一个小的技术细节，我们需要指定扰动样式变量 $\widetilde{\mathbf{s}}_1$ 和 $\widetilde{\mathbf{s}}_2$ 的条件生成，以确保它们在相对于 s 的扰动方式上并不完全相同。这可以通过适当约束 p_A 来确保，以排除 $\widetilde{\mathbf{s}}_1$ 和 $\widetilde{\mathbf{s}}_2$ 中维度完全对齐的退化情况——这是在定理 1 的证明第 2 步中需要排除的情况。

B EXPERIMENTS

B 实验

B.1 EXPERIMENTAL DETAILS

B.1 实验细节

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 (Hyvärinen 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. In Table 2a, we specify the main hyperparameters for the numerical simulation.

数值仿真生成过程在第 5.1 节中描述。这里，我们提供有关实验的额外信息。可逆 MLP 的构建类似于之前的工作 (Hyvärinen 和 Morioka, 2016; Hyvärinen 和 Morioka, 2017; Zimmermann 等, 2021; von Kügelgen 等, 2021)，通过重新抽样平方权重矩阵，直到其条件数超过阈值。在原始设置 ($\mathbf{f}_1 = \mathbf{f}_2$) 中，我们使用一个编码器 ($\mathbf{g}_1 = \mathbf{g}_2$) ，而在多模态设置 ($\mathbf{f}_1 \neq \mathbf{f}_2$) 中，我们使用不同的编码器 ($\mathbf{g}_1 \neq \mathbf{g}_2$) 来反映不同混合函数的假设，并且在实践中，观察的维度在不同模态之间可能会有所不同。在表 2a 中，我们指定了数值仿真的主要超参数。

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 [7] 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 $[0, 1]$ 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 $[0, 1/3]$. 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. In Appendix B.2, we provide additional results for a version of the dataset with mutually independent factors.

Multimodal3DIdent 我们的图像/文本对数据集基于用于生成 Causal3DIdent (von Kügelgen et al., 2021; Zimmermann et al., 2021) 和 CLEVR (Johnson et al., 2017) 数据集的代码。图像是使用 Blender 渲染器生成的 (Blender Online Community, 2018)。渲染作为一个复杂的混合函数，从表 3 中列出的 11 个不同参数 (即潜在因素) 生成图像。为了生成文本描述，我们改编了来自 CLEVR (Johnson et al., 2017) 的文本渲染，并使用 5 种不同的短语来引导特定于模态的变化。用于生成文本的潜在因素也列在表 3 中。图像和文本模态之间的依赖关系由三个内容因素 (物体形状、x -位置和 y -位置) 和一个风格因素 (物体颜色) 决定。对于图像中的物体颜色，我们使用连续的色调值，而对于文本，我们将 RGB 值与三个不同调色板 [7] 中的最近颜色值匹配，该调色板在每次观察中均匀随机抽样。此外，我们确保在调色板之间没有重叠的颜色值，必要时为各自的调色板使用前缀 (例如，"xkcd:black")。在第 5.2 节中，我们使用一个版本的 Multimodal3DIdent 数据集，该数据集具有从内容到风格的因果依赖关系。具体而言，物体的颜色依赖于其 x 位置。特别地，我们将色调值范围 $[0, 1]$ 分成三个大小相等的区间，并将每个区间与物体的固定 x 位置关联。例如，如果 x -位置为 "左"，我们从区间 $[0, 1/3]$ 中抽样色调值。因此，物体的颜色可以在一定程度上根据物体的位置进行预测。来自 Multimodal3DIdent 数据集的图像/文本对样本如图 2 和图 4 所示。实验的超参数列在表 2b 中。在附录 B.2 中，我们提供了一个具有相互独立因素的数据集版本的额外结果。

High-dimensional image pairs In Appendix B. 2, we provide additional results using a dataset of high-dimensional pairs of images of size $224 \times 224 \times 3$. Similar to Multimodal3DIdent, images are generated using Blender (Blender Online Community, 2018) and code adapted from previous work (Zimmermann et al., 2021; von Kügelgen et al., 2021). Each image depicts a scene with one type of object (a teapot, like in Zimmermann et al., 2021) in front of a colored background and illuminated by a colored spotlight (for examples, see Figure 9). The scene is defined by 9 continuous latent variables each of which is sampled from a uniform distribution. Object positions (x-, y- and z-coordinates) are content factors that are always shared between modalities, while object-, spotlight-and background-colors are style factors that are stochastically shared. Modality-specific factors are object rotation (α and β angles) for one modality and spotlight position for the other. To simulate modality-specific mixing functions, we render the objects using distinct textures (i.e., rubber and metallic) for each modality. Further, we generate two versions of this dataset, with and without

高维图像对在附录 B.2 中，我们提供了使用高维图像对数据集的额外结果，图像大小为 $224 \times 224 \times 3$ 。与 Multimodal3DIdent 类似，图像是使用 Blender(Blender 在线社区，2018) 生成的，并且代码改编自之前的工作 (Zimmermann 等，2021；von Kügelgen 等，2021)。每幅图像描绘了一个场景，场景中有一种类型的物体 (如 Zimmermann 等，2021 中的茶壶)，背景为彩色，并由彩色聚光灯照明 (例如，见图 9)。场景由 9 个连续的潜变量定义，每个变量均从均匀分布中抽样。物体位置 (x、y 和 z 坐标) 是始终在模态之间共享的内容因子，而物体、聚光灯和背景颜色是随机共享的风格因子。特定于模态的因子是一个模态的物体旋转 (α 和 β 角度) 和另一个模态的聚光灯位置。为了模拟特定于模态的混合函数，我们为每个模态使用不同的纹理 (即，橡胶和金属) 来渲染物体。此外，我们生成了该数据集的两个版本，有因果依赖和没有因果依赖。

causal dependencies. For the dataset with causal dependencies we sample the latent factors according to a causal model, where background-color depends on z-position and spotlight-color depends on object-color. We use ResNet-18 encoders and similar hyperparameter values to those used for the image/text experiment (Table 2b).

对于具有因果依赖的数据集，我们根据因果模型抽样潜因子，其中背景颜色依赖于 z 位置，聚光灯颜色依赖于物体颜色。我们使用 ResNet-18 编码器和与图像/文本实验 (表 2b) 中使用的超参数值相似的超参数值。

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

参数	值
生成函数	3 层多层感知器 (MLP)
编码器	7 层多层感知器 (MLP)
优化器	Adam
条件阈值比	1e-3
维度 d	15
批量大小	6144
学习率	1e-4
温度 τ	1.0
# 种子	3
# 迭代	300,000
相似性度量	欧几里得
梯度裁剪	2-范数；最大值 2

(a) Parameters used for the numerical simulation.
(a) 用于数值模拟的参数。

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 2

参数	值
生成函数	图像和文本渲染
图像编码器	ResNet-18
文本编码器	4 层卷积网络
优化器	Adam
批量大小	256
学习率	1e-5
温度 τ	1.0
# 种子	3
# 迭代	100,000
# 样本 (训练 / 验证 / 测试)	125,000/10,000/10,000
相似度度量	余弦相似度
梯度裁剪	2-范数；最大值 2

(b) Parameters used for Multimodal3DIdent.

(b) 用于 Multimodal3DIdent 的参数。

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

表 2: 在主文本中用于两个实验的实验参数和超参数。

Latent factor	Distribution	Details
Object shape	Categorical	7 unique values
Object x-position	Categorical	3 unique values
Object y-position	Categorical	3 unique values
Object color	Uniform	hue value in $[0, 1]$
Object rotation α	Uniform	angle value in $[0, 1]$
Object rotation β	Uniform	angle value in $[0, 1]$
Object rotation γ	Uniform	angle value in $[0, 1]$
Spotlight position	Uniform	angle value in $[0, 1]$
Spotlight color	Uniform	hue value in $[0, 1]$
Background color	Uniform	hue value in $[0, 1]$
Object shape	Categorical	7 unique values
Object x-position	Categorical	3 unique values
Object y-position	Categorical	3 unique values
Object color	Categorical	color names (3 palettes)
Text phrasing	Categorical	5 unique values

潜在因子	分布	细节
物体形状	分类的	7 个独特值
对象 x 位置	分类的	3 个独特值
对象 y 位置	分类的	3 个独特值
对象颜色	均匀	$[0, 1]$ 中的色调值
对象旋转 α	均匀	$[0, 1]$ 中的角度值
对象旋转 β	均匀	$[0, 1]$ 中的角度值
对象旋转 γ	均匀	$[0, 1]$ 中的角度值
聚光灯位置	均匀	$[0, 1]$ 中的角度值
聚光灯颜色	均匀	$[0, 1]$ 中的色调值
背景颜色	均匀	$[0, 1]$ 中的色调值
物体形状	分类的	7 个独特值
对象 x 位置	分类的	3 个独特值
对象 y 位置	分类的	3 个独特值
对象颜色	分类的	颜色名称 (3 种调色板)
文本措辞	分类的	5 个独特值

Table 3: Description of the latent factors used to generate Multimodal3DIdent. The first 10 factors are used to generate the images and the remaining 5 factors are used to generate the text. Object z-position is kept constant for all images, which is why we do not list it among the generative factors. Independent factors are drawn uniformly from the respective distribution. Content factors are denoted in bold and style factors in italic; the remaining factors are modality-specific.

表 3: 用于生成 Multimodal3DIdent 的潜在因素描述。前 10 个因素用于生成图像，其余 5 个因素用于生成文本。所有图像的对象 z 位置保持不变，这就是我们不将其列为生成因素的原因。独立因素均匀地从各自的分布中抽取。内容因素用粗体表示，风格因素用斜体表示；其余因素是特定于模态的。

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

[7] 我们使用来自 matplotlib.colors API 的以下三种调色板:Tableau 颜色 (10 个值)、CSS4 颜色 (148 个值) 和 XKCD 颜色 (949 个值)。

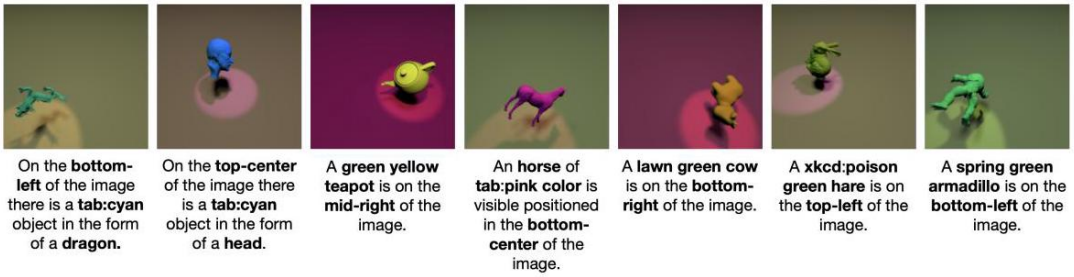


Figure 4: Examples of image/text pairs from the Multimodal3DIdent dataset. Each sample shows one of the seven shapes or classes of objects included in the dataset.

图 4: 来自 Multimodal3DIdent 数据集的图像/文本对示例。每个样本展示了数据集中包含的七种形状或类别之一。

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	X	X	1.00 ± 0.00	0.00 ± 0.00
0.75	x	X	0.99 ± 0.01	0.00 ± 0.00
0.75	✓	X	0.99 ± 0.00	0.52 ± 0.09
0.75	X	$\surd$	1.00 ± 0.00	0.79 ± 0.04
0.75	✓	3	0.99 ± 0.01	0.81 ± 0.04

生成过程			R^2 (非线性)	
p(chg.)	统计	因果	内容 c	风格 s
1.0	X	X	1.00 ± 0.00	0.00 ± 0.00
0.75	x	X	0.99 ± 0.01	0.00 ± 0.00
0.75	✓	X	0.99 ± 0.00	0.52 ± 0.09
0.75	X	$\surd$	1.00 ± 0.00	0.79 ± 0.04
0.75	✓	3	0.99 ± 0.01	0.81 ± 0.04

(a) Original setting

(a) 原始设置

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	x	x	1.00 ± 0.00	0.00 ± 0.00
0.75	x	x	$\mathbf{1.00} \pm 0.00$	0.00 ± 0.00
0.75	3	x	0.99 ± 0.01	0.36 ± 0.10
0.75	X	✓	$\mathbf{1.00} \pm 0.00$	0.81 ± 0.03
0.75	$\surd$	✓	0.99 ± 0.01	$\mathbf{0.83} \pm 0.05$

生成过程			R^2 (非线性)	
p(变化)	统计	因果	内容 c	风格 s
1.0	x	x	1.00 ± 0.00	0.00 ± 0.00
0.75	x	x	$\mathbf{1.00} \pm 0.00$	0.00 ± 0.00
0.75	3	x	0.99 ± 0.01	0.36 ± 0.10
0.75	X	✓	$\mathbf{1.00} \pm 0.00$	0.81 ± 0.03
0.75	$\surd$	✓	0.99 ± 0.01	$\mathbf{0.83} \pm 0.05$

(b) Multimodal setting

(b) 多模态设置

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.

表 4: 没有特定于模态的潜在变量的数值模拟结果。我们将原始设置 ($\mathbf{f}_1 = \mathbf{f}_2$ ，左侧表格) 与多模态设置 ($\mathbf{f}_1 \neq \mathbf{f}_2$ ，右侧表格) 进行比较。每一行展示了不同设置的结果，具有不同的风格变化概率 p(chg.) 和可能的统计 (Stat.) 和/或因果 (Caus.) 依赖关系。每个值表示非线性回归模型的 R^2 决定系数 (在 3 个种子上取平均)，该模型预测相应的真实因素 ($\mathbf{c}, \mathbf{s}$ 或 $\mathbf{m}_i$) 。

B.2 ADDITIONAL EXPERIMENTAL RESULTS

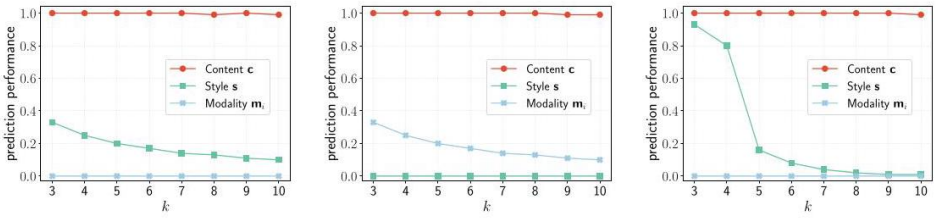
B.2 额外实验结果

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.

无模态特定潜变量的数值模拟回顾所考虑的生成过程 (第 3 节) 有两个模态特定变异的来源: 模态特定混合函数和模态特定潜变量。为了去耦这两种变异来源的影响，我们进行了不使用模态特定潜变量的消融研究。表 4 展示了结果，表明在原始设置 ($\mathbf{f}_1 = \mathbf{f}_2$ ，表 4a) 和多模态设置 ($\mathbf{f}_1 \neq \mathbf{f}_2$ ，表 4b) 中内容是块识别的。与表 1 相比，我们观察到在没有模态特定潜变量的情况下，内容预测略有改善。因此，我们的结果表明，对比学习可以在有和没有模态特定潜变量的多模态设置中块识别内容因素。

Numerical simulation with discrete latent factors Extending the numerical simulation from Section 5.1, we test block-identifiability of content information when observations are generated from a mixture of continuous and discrete latent variables, thus violating one of the assumptions from Theorem 1. In this setting, content, style and modality-specific information are random variables with 5 components sampled from either a continuous normal distribution or a discrete multinomial distribution with k classes, for which we experiment with different $k \in \{3, 4, \ldots, 10\}$. For all settings, we train an encoder with the InfoNCE objective and set the encoding size to 5 dimensions. The other hyperparameters used in this set of experiments are detailed in Table 2a. To ensure convergence of the models, we extended the number of training iterations to 600,000 and 3,000,000 for experiments with discrete style/modality-specific and discrete content variables respectively. With discrete style or modality-specific variables and continuous content (Figures 5a and 5b), the results suggest that content is block-identified, since the prediction of style and modality-specific information is at chance level (i.e., accuracy $= 1/k$) while content is consistently fully recovered $\left(R^2 \geq 0.99\right)$. In the opposite setting, with continuous style and modality-specific variables and discrete content (Figure 5c), the number of content classes appears to be a critical factor for block-identifiability of content: while content is always encoded well, style information is also encoded to a significant extent when the number of content classes is small, but significantly less style can be recovered when the number

带有离散潜在因子的数值模拟扩展第 5.1 节中的数值模拟，我们测试当观察值是从连续和离散潜在变量的混合中生成时内容信息的块可识别性，从而违反了定理 1 中的一个假设。在这种情况下，内容、风格和特定模态的信息是具有 5 个成分的随机变量，这些成分是从连续正态分布或具有 k 类的离散多项分布中抽样的，我们对不同的 $k \in \{3, 4, \ldots, 10\}$ 进行了实验。在所有设置中，我们使用 InfoNCE 目标训练编码器，并将编码大小设置为 5 维。该组实验中使用的其他超参数详见表 2a。为了确保模型的收敛性，我们将训练迭代次数扩展到 600,000 次和 3,000,000 次，分别用于离散风格/模态特定和离散内容变量的实验。对于离散风格或模态特定变量和连续内容 (图 5a 和 5b)，结果表明内容是块可识别的，因为风格和模态特定信息的预测处于偶然水平 (即，准确率 $= 1/k$)，而内容始终被完全恢复 $\left(R^2 \geq 0.99\right)$ 。在相反的情况下，对于连续风格和模态特定变量以及离散内容 (图 5c)，内容类别的数量似乎是内容块可识别性的关键因素: 虽然内容总是被很好地编码，但当内容类别数量较少时，风格信息也被相当程度地编码，但当内容类别数量增加时，恢复的风格信息显著减少。



(a) Only $\mathbf{s}$ is discrete (b) Only $\mathbf{m}_1, \mathbf{m}_2$ are discrete (c) Only $\mathbf{c}$ is discrete

Figure 5: Numerical simulations with discrete latent factors. The results show three settings in each of which one group of latent variables is discrete while the remaining groups are continuous. Continuous variables are normally distributed, whereas discrete variables are sampled from a multinomial distribution with k distinct classes. We measure the prediction performance with a nonlinear model 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 error bars show the standard deviation, which is relatively small.

图 5: 具有离散潜在因素的数值模拟。结果显示了三种设置，其中一组潜在变量是离散的，而其余组是连续的。连续变量服从正态分布，而离散变量则是从具有 k 个不同类别的多项分布中抽样得到的。我们用非线性模型测量预测性能，分别以连续因素的 R^2 决定系数和离散因素的分类准确率为标准。每个点表示三个种子的平均值，误差条表示标准差，相对较小。

of content classes increases. Through this set of experiments, we challenge the assumption that all generative factors should be continuous (c.f., Section 3) and show that block-identifiability of content can still be satisfied when content is continuous while style or modality-specific variables are discrete. On the other hand, style is encoded to a significant extent when content is discrete, which might explain our observation for the image/text experiment, where we saw that, in the presence of discrete content factors, some style information can be encoded.

内容类别的数量增加。通过这一系列实验，我们挑战了所有生成因素应为连续的假设 (参见第 3 节)，并展示了当内容是连续的而风格或模态特定变量是离散时，内容的块可识别性仍然可以得到满足。另一方面，当内容是离散时，风格被显著编码，这可能解释了我们在图像/文本实验中的观察结果，在该实验中，我们看到在存在离散内容因素的情况下，一些风格信息可以被编码。

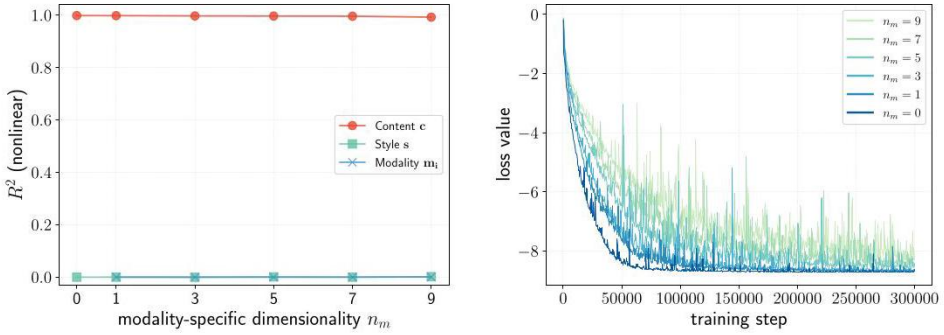
Dimensionality ablations for the numerical simulation To test the effect of latent dimensionality on identifiability, Figure 6 presents dimensionality ablations where we keep the number of content dimensions fixed and only vary the number of style or modality-specific dimensions, n_s and n_m respectively. Figures 6a and 6c confirm that block-identifiability of content still holds when we significantly increase the number of style or modality-specific dimensions, as the representation consistently encodes only content and no style or modality-specific information. In Figures 6b and 6d, we can observe that the training loss decreases more slowly when we increase the dimensionality of n_c and n_s respectively, which provides an intuition that the sample complexity might increase with the number of style and modality-specific dimensions.

数值模拟的维度消融实验为了测试潜在维度对可识别性的影响，图 6 展示了维度消融实验，其中我们固定内容维度的数量，仅改变风格或模态特定维度的数量，n_s 和 n_m 。图 6a 和 6c 确认，当我们显著增加风格或模态特定维度的数量时，内容的块可识别性仍然成立，因为表示始终仅编码内容，而没有风格或模态特定信息。在图 6b 和 6d 中，我们可以观察到，当我们分别增加 n_c 和 n_s 的维度时，训练损失下降得更慢，这提供了一个直觉，即样本复杂性可能随着风格和模态特定维度的数量增加而增加。

Estimating the number of content factors The estimation of the number of content factors is an important puzzle piece, since Theorem 1 assumes that the number of content factors is known or that it can be estimated. In practice, the number of content factors can be viewed as a single hyperparameter (e.g., Locatello et al., 2020) that can be tuned with respect to a suitable model selection metric. For instance, one could use the validation loss for model selection, which would be convenient since the validation loss only requires a holdout dataset and no additional supervision. In Figure 7, we plot the validation loss (averaged over 2,000 validation samples) as a function of the encoding size for both experiments used in our paper. Results for the numerical simulation are shown in Figure 7a and for the image/text experiment in Figure 7b. For both datasets, we observe that the validation loss increases most significantly in the range around the true number of content factors. For the numerical simulation, the results look promising as they show a clear "elbow" (James et al., 2013) at the correct value of 5, which corresponds to the true number of content factors. The results are less clear for the image/text experiment, where the elbow method might suggest the range of 2-4 content factors, while the true value

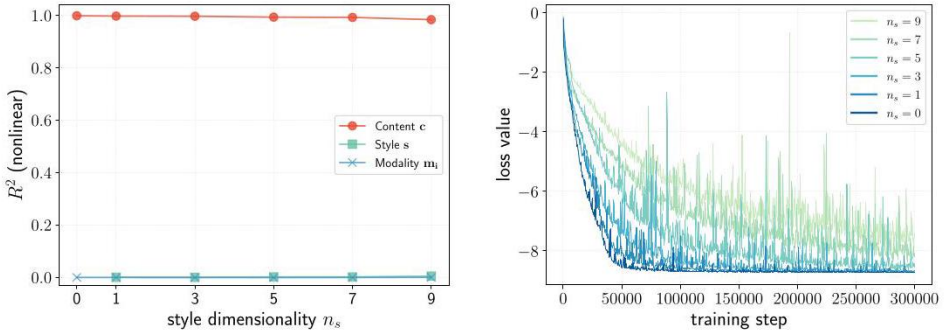
is 3 . While these initial results look promising, we believe that more work is required to investigate the estimation of the number of content factors and the design of suitable heuristics, which are interesting directions for future research.

估计内容因子的数量内容因子数量的估计是一个重要的难题，因为定理 1 假设内容因子的数量是已知的或可以被估计。在实践中，内容因子的数量可以视为一个单一的超参数 (例如，Locatello 等，2020)，可以根据适当的模型选择指标进行调整。例如，可以使用验证损失进行模型选择，这将是方便的，因为验证损失只需要一个保留数据集而不需要额外的监督。在图 7 中，我们绘制了验证损失 (在 2000 个验证样本上取平均) 与编码大小的关系，针对我们论文中使用的两个实验。数值模拟的结果显示在图 7a 中，图像/文本实验的结果显示在图 7b 中。对于这两个数据集，我们观察到验证损失在真实内容因子数量附近的范围内显著增加。对于数值模拟，结果看起来很有前景，因为它们在正确的值 5 处显示出明显的 “肘部” (James 等，2013)，这对应于真实的内容因子数量。对于图像/文本实验，结果则不太明确，肘部法可能建议 2-4 个内容因子的范围，而真实值为 3。尽管这些初步结果看起来很有前景，但我们认为需要更多的工作来研究内容因子数量的估计以及设计合适的启发式方法，这些都是未来研究的有趣方向。



(a) Prediction performance as a function of n_m

(b) Learning curves for different n_m

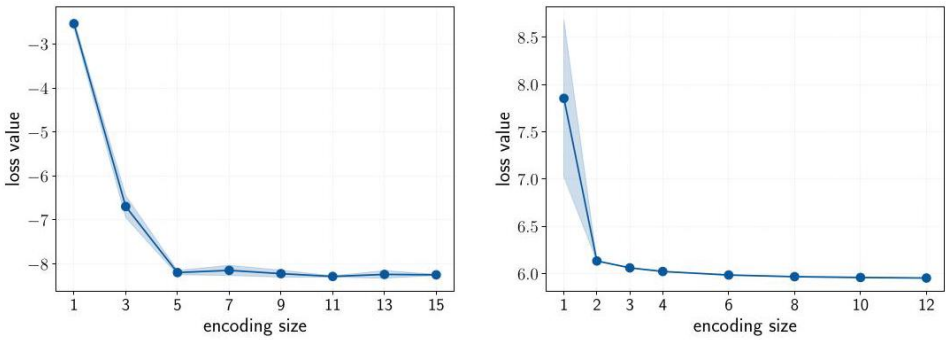


(c) Prediction performance as a function of n_s

(d) Learning curves for different n_s

Figure 6: Dimensionality ablation for the numerical simulation. We consider the multimodal setting with mutually independent factors and test the effect of latent dimensionality on identifiability by keeping the number of content dimensions fixed and only varying the number of style or modality-specific dimensions (n_s and n_m respectively). In Figures 6a and 6c we measure the nonlinear prediction performance in terms of the R^2 coefficient of determination of a nonlinear regression model that predicts the respective ground truth factor ($\mathbf{c}, \mathbf{s}$, or $\mathbf{m}_i$) from the learned representation. In Figures 6 b and 6 d , we plot the learning curves (i.e., the training loss) of the respective models to compare how fast they converge.

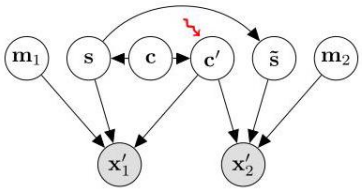
图 6: 数值仿真的维度消融。我们考虑多模态设置，其中因素相互独立，并通过保持内容维度的数量固定，仅改变样式或特定模态维度的数量来测试潜在维度对可识别性的影响 (分别为 n_s 和 n_m)。在图 6a 和 6c 中，我们测量非线性回归模型的非线性预测性能，以确定系数 R^2 ，该模型预测从学习到的表示中获得的相应真实因素 ($\mathbf{c}, \mathbf{s}$, 或 $\mathbf{m}_i$) 。在图 6 b 和 6 d 中，我们绘制了各自模型的学习曲线 (即训练损失)，以比较它们收敛的速度。



(a) Validation loss for the numerical simulation (b) Validation loss for the image/text experiment

Figure 7: An attempt at estimating of the number of content factors using the validation loss. The validation loss corresponds to the value of the $\mathcal{L}_{\text{SymInfoNCE}}$ objective computed on a holdout dataset. Since we are interested in estimating the true number of content factors to select the encoding size appropriately, we plot the validation loss as a function of the encoding size. We show the validation loss for the numerical simulation with independent factors (Figure 7a) and for the image/text experiment (Figure 7b) respectively.

图 7: 尝试使用验证损失估计内容因素的数量。验证损失对应于在保留数据集上计算的 $\mathcal{L}_{\text{SymInfoNCE}}$ 目标值。由于我们希望估计真实的内容因素数量以适当地选择编码大小，因此我们将验证损失绘制为编码大小的函数。我们分别展示了独立因素的数值仿真 (图 7a) 和图像/文本实验 (图 7b) 的验证损失。



Generative process			R^2 (nonlinear)			
p(chg.)	Stat.	Cau.	Content c	Content c'	Style s	Modality m_i
1.0	x	x	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	x	x	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	✓	x	0.00 ± 0.00	$\mathbf{1.00} \pm 0.00$	0.50 ± 0.19	0.00 ± 0.00
0.75	x	✓	0.01 ± 0.00	$\mathbf{0.98} \pm 0.00$	0.03 ± 0.01	0.00 ± 0.00
0.75	✓	✓	0.28 ± 0.14	$\mathbf{0.91} \pm 0.03$	$\underline{0.39} \pm 0.20$	0.00 ± 0.00

生成过程			R^2 (非线性)			
p(chg.)	统计	因果	内容 c	内容 c'	风格 s	模态 m_i
1.0	x	x	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	x	x	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
0.75	✓	x	0.00 ± 0.00	$\mathbf{1.00} \pm 0.00$	0.50 ± 0.19	0.00 ± 0.00
0.75	x	✓	0.01 ± 0.00	$\mathbf{0.98} \pm 0.00$	0.03 ± 0.01	0.00 ± 0.00
0.75	✓	✓	0.28 ± 0.14	$\mathbf{0.91} \pm 0.03$	$\underline{0.39} \pm 0.20$	0.00 ± 0.00

Figure 8: Evaluation with test-time interventions. We use the interventional setup that is illustrated on the left, i.e., perturbed samples $\mathbf{x}'_1, \mathbf{x}'_2$ that are generated from the intervened content $\mathbf{c}'$, which is a copy of the original content $\mathbf{c}$ with an intervention, i.e., a batch-wise permutation ($\rightsquigarrow$) that makes $\mathbf{c}'$ independent of $\mathbf{s}$. 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 ($\mathbf{c}, \mathbf{c}', \mathbf{s}$, or $\mathbf{m}_i$) from the learned representation.

图 8: 测试时干预的评估。我们使用左侧所示的干预设置，即，从干预内容 $\mathbf{c}'$ 生成的扰动样本 $\mathbf{x}'_1, \mathbf{x}'_2$ ，该内容是原始内容 $\mathbf{c}$ 的副本，并进行了干预，即，批量置换 ($\rightsquigarrow$) ，使得 $\mathbf{c}'$ 与 $\mathbf{s}$ 独立。每一行展示了不同设置的结果，具有不同的风格变化概率 p(chg.) 和可能的统计 (Stat.) 和/或因果 (Caus.) 依赖关系。

每个值表示 R^2 决定系数 (在 3 个种子上平均) 用于预测相应的真实因素 ($\mathbf{c}, \mathbf{c}', \mathbf{s}$, or $\mathbf{m}_i$) 的非线性回归模型的结果。

Evaluation with test-time interventions Previously, we observed that style can be predicted to some degree when there are causal dependencies from content to style (Table 1), which can be attributed to style information being partially predictable from the encoded content information in the causal setup. To verify that the encoders only depend on content information (i.e., that content is block-identified), we assess the trained models using a novel, more rigorous empirical evaluation for the numerical simulation. We test the effect of interventions $\mathbf{c} \to \mathbf{c}'$, which perturb the content information at test time via batch-wise permutations of content, before generating $\mathbf{x}'_1 = \mathbf{f}_1(\mathbf{c}', \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{x}'_2 = \mathbf{f}_1(\mathbf{c}', \widetilde{\mathbf{s}}, \mathbf{m}_1)$. Hence, we break the causal dependence between content and style (see illustration in Figure 8), which allows us to better assess whether the trained encoders depend on content or style information. Specifically, we train the encoders for 3,000,000 iterations to ensure convergence and then train nonlinear regression models to predict both the original and the intervened content variables from the learned representations. Figure 8 presents our results using the interventional setup, showing that in most cases only content information can be recovered. We observe an exception (underlined values) in the two cases with statistical dependencies, where some style information can be recovered, which is expected because statistical dependencies reduce the effective dimensionality of content (cp. von Kügelgen et al., 2021). Analogously, in the case of statistical and causal dependencies, some of the original content information can be recovered via the encoded style information. In summary, the evaluation with interventions provides a more rigorous assessment of block-identifiability in the causal setup, showing that neither style nor modality-specific information can be recovered when the encoding size matches the true number of content dimensions.

测试时干预的评估之前，我们观察到，当内容与风格之间存在因果依赖关系时，风格在某种程度上是可以被预测的 (见表 1)，这可以归因于风格信息在因果设置中部分可从编码的内容信息中预测。为了验证编码器仅依赖于内容信息 (即内容是块识别的)，我们使用一种新颖且更严格的实证评估方法来评估训练好的模型，以进行数值模拟。我们测试干预的效果 $\mathbf{c} \to \mathbf{c}'$ ，这些干预通过批量置换内容在测试时扰动内容信息，然后生成 $\mathbf{x}'_1 = \mathbf{f}_1(\mathbf{c}', \mathbf{s}, \mathbf{m}_1)$ 和 $\mathbf{x}'_2 = \mathbf{f}_1(\mathbf{c}', \widetilde{\mathbf{s}}, \mathbf{m}_1)$ 。因此，我们打破了内容与风格之间的因果依赖 (见图 8 中的插图)，这使我们能够更好地评估训练好的编码器是否依赖于内容或风格信息。具体而言，我们训练编码器 3000000 次以确保收敛，然后训练非线性回归模型，以从学习到的表示中预测原始和干预后的内容变量。图 8 展示了我们使用干预设置的结果，表明在大多数情况下仅能恢复内容信息。我们观察到在两个具有统计依赖关系的案例中存在例外 (下划线值)，在这些情况下可以恢复一些风格信息，这是可以预期的，因为统计依赖关系降低了内容的有效维度 (参见 von Kügelgen 等，2021)。类似地，在统计和因果依赖的情况下，部分原始内容信息可以通过编码的风格信息恢复。总之，干预评估提供了对因果设置中块可识别性的更严格评估，表明当编码大小与内容维度的真实数量匹配时，既无法恢复风格信息，也无法恢复特定于模态的信息。

High-dimensional image pairs with continuous latents To bridge the gap between continuous and discrete data, we provide an additional experiment that offers a realistic setup but uses only continuous latent variables to satisfy the assumptions of Theorem 1. Previously, in Section 5.2, we used a complex multimodal dataset of image/text pairs, which were generated from a combination of continuous and discrete latent factors. Now, we consider a different dataset that consists of pairs high-dimensional images generated from a set of continuous latents, which is more in line with our theoretical assumptions. Note that datasets with pairs of images are common in practice, for example, in medical imaging where patients are assessed using multiple views (e.g., images from different angles) or multiple modalities (e.g., as in PET-CT imaging). To generate the data, we adapt the code from 3DIdent (Zimmermann et al., 2021) to render pairs of images, for which the object position is always shared (i.e., content), the object-, spotlight- and background-color is stochastically shared (i.e., style), and modality-specific factors are object rotation for one modality and spotlight position for the other. Additionally, we render the objects using different textures to simulate a modality-specific mixing process. Samples of image pairs are shown in Figure 9 and further details about the dataset can be found in Appendix B.1. We train the encoders with the InfoNCE objective for 60,000 iterations using the same architectures and hyperparameters as for Multimodal3DIdent (Table 2b), and again evaluate the R^2 coefficient of determination using a kernel ridge regression that predicts the respective ground truth factor from the learned representations.

高维图像对与连续潜变量为了弥合连续数据和离散数据之间的差距，我们提供了一个额外的实验，该实验提供了一个现实的设置，但仅使用连续潜变量以满足定理 1 的假设。之前在第 5.2 节中，我们使用了一个复杂的多模态数据集，该数据集由图像/文本对组成，这些对是由连续和离散潜因子的组合生成的。现在，我们考虑一个不同的数据集，该数据集由一组连续潜变量生成的高维图像对组成，这更符合我们的理论假设。请注意，具有图像对的数据集在实践中是常见的，例如，在医学成像中，患者使用多个视图 (例如，从不同角度的图像) 或多种模态 (例如，PET-CT 成像) 进行评估。为了生成数据，我们改编了 3DIdent(Zimmermann 等，2021) 的代码，以渲染图像对，其中对象位置始终是共享的 (即内容)，

对象、聚光灯和背景颜色是随机共享的 (即风格)，而特定模态的因素是一个模态的对象旋转和另一个模态的聚光灯位置。此外，我们使用不同的纹理渲染对象，以模拟特定模态的混合过程。图像对的样本如图 9 所示，关于数据集的更多细节可以在附录 B.1 中找到。我们使用与 Multimodal3DIdent(表 2b) 相同的架构和超参数，以 InfoNCE 目标训练编码器 60,000 次迭代，并再次使用核岭回归评估 R^2 决定系数，该回归预测从学习到的表示中得到的相应真实因子。

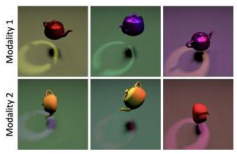
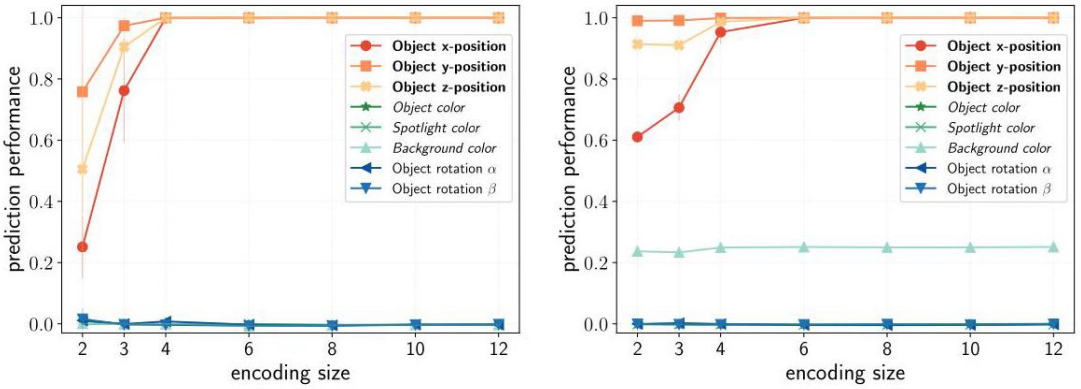


Figure 9: Examples of high-dimensional image pairs.
图 9: 高维图像对的示例。



(a) With mutually independent factors

(b) With causal dependencies

Figure 10: Result with pairs of high-dimensional images. As a function of the encoding size of the model, we assess the nonlinear prediction of ground truth factors to quantify how well the learned representation encodes the respective factors. Content factors are denoted in bold, style factors in italic, and modality-specific factors in regular font. Each point denotes the average R^2 score across three seeds and bands show one standard deviation.

图 10: 高维图像对的结果。作为模型编码大小的函数，我们评估了对真实因素的非线性预测，以量化学习到的表示如何编码各自的因素。内容因素用粗体表示，风格因素用斜体表示，特定模态的因素用常规字体表示。每个点表示三个种子之间的平均 R^2 分数，带状图显示一个标准差。

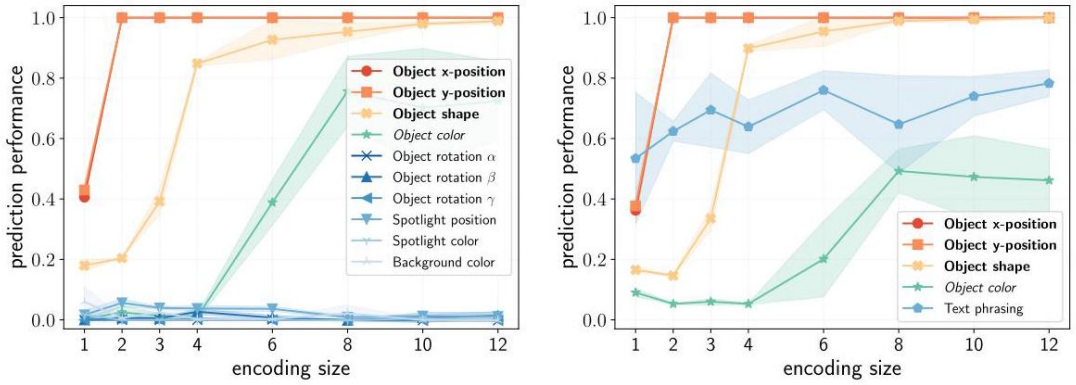
Figure 10 present our results for the dataset of image pairs, showing the prediction performance as a function of the encoding size for the setting with causal dependencies (Figure 10b) and the setting with mutually independent latent variables (Figure 10a) respectively. In both settings, content information (i.e. object position) is recovered when sufficient encoding capacity is available. Style and modality-specific information, on the other hand, are discarded independent of the encoding size. In Figure 10b we observe the recovery of some style information, which is expected because style can be predicted to some degree from the encoded content information when there is a causal dependence of style on content. Overall, these findings lend further support to our theoretical result from Theorem 1, as we investigate a realistic setting with only continuous latent factors, which is more in line with our assumptions. Notably, the results appear more consistent with our theory, e.g., showing that less style and modality-specific information is encoded, compared to our results for the image/text experiment, where we used a combination of continuous and discrete latent factors.

图 10 展示了图像对数据集的结果，分别显示了在具有因果依赖关系的设置 (图 10b) 和具有相互独立的潜变量的设置 (图 10a) 下，预测性能作为编码大小的函数。在这两种设置中，当编码容量足够时，内容信息 (即对象位置) 得以恢复。另一方面，风格和特定模态的信息则独立于编码大小而被丢弃。在图 10b 中，我们观察到一些风格信息的恢复，这是可以预期的，因为当风格依赖于内容时，风格可以在一定程度上从编码的内容信息中预测。总体而言，这些发现进一步支持了我们从定理 1 得出的理论结果，因为我们研究了仅具有连续潜因素的现实设置，这与我们的假设更为一致。值得注意的是，结果似乎与我

们的理论更加一致，例如，显示出编码的风格和特定模态的信息较少，与我们在图像/文本实验中的结果相比，在该实验中我们使用了连续和离散潜因素的组合。

Multimodal3DIdent with mutually independent factors For the results of the image/text experiment in the main text (Section 5.2) we used the Multimodal3DIdent dataset, which we designed such that object color is causally dependent on the x-position of the object to impose a causal dependence of style on content. In Figure 11, we provide a similar analysis using a version of the dataset without the causal dependence, i.e., with mutually independent factors. For both modalities, we observe that object color is only encoded when the encoding size is larger than four, i.e., when there is excess capacity beyond the capacity needed to encode all content factors. Hence, these results corroborate that contrastive learning can block-identify content factors in a complex multimodal setting with

Multimodal3DIdent 具有相互独立的因素。对于主文中图像/文本实验的结果 (第 5.2 节)，我们使用了 Multimodal3DIdent 数据集，该数据集的设计使得对象颜色在因果上依赖于对象的 x 位置，以施加风格对内容的因果依赖。在图 11 中，我们提供了使用没有因果依赖的数据集版本的类似分析，即具有相互独立因素的版本。对于这两种模态，我们观察到只有在编码大小大于四时，对象颜色才会被编码，即当存在超出编码所有内容因素所需的容量时。因此，这些结果证实了对比学习能够在复杂的多模态环境中阻止识别内容因素。



(a) Prediction of image factors

(b) Prediction of text factors

Figure 11: Result on Multimodal3DIdent with mutually independent factors. 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.

图 11: 在具有相互独立因素的 Multimodal3DIdent 上的结果。作为模型编码大小的函数，我们评估真实图像因素 (左子图) 和文本因素 (右子图) 的非线性预测，以量化学习到的表示如何编码各自的因素。内容因素用粗体表示，风格因素用斜体表示。沿 x 轴，我们改变编码大小，即模型的输出维度。我们分别以连续因素的 R^2 决定系数和离散因素的分类准确度来衡量预测性能。每个点表示三个种子的平均值，带状图显示一个标准差。