NEGATE OR EMBRACE: ON HOW MISALIGNMENT SHAPES MULTIMODAL REPRESENTATION LEARNING

Yichao Cai[†] **Yuhang Liu**[†] **Erdun Gao** **Tianjiao Jiang** **Zhen Zhang**

Anton van den Hengel **Javen Qinfeng Shi**

Australian Institute for Machine Learning
The University of Adelaide, SA 5000, Australia
`https://github.com/YichaoCai1/crossmodal_mislaignment`

ABSTRACT

Multimodal representation learning, exemplified by multimodal contrastive learning (MMCL) using image-text pairs, aims to learn powerful representations by aligning cues across modalities. This approach relies on the core assumption that the exemplar image-text pairs constitute two representations of an identical concept. However, recent research has revealed that real-world datasets often exhibit *misalignment*. There are two distinct viewpoints on how to address this issue: one suggests mitigating the misalignment, and the other leveraging it. We seek here to reconcile these seemingly opposing perspectives, and to provide a practical guide for practitioners. Using latent variable models we thus formalize misalignment by introducing two specific mechanisms: *selection bias*, where some semantic variables are missing, and *perturbation bias*, where semantic variables are distorted—both affecting latent variables shared across modalities. Our theoretical analysis demonstrates that, under mild assumptions, the representations learned by MMCL capture exactly the information related to the subset of the semantic variables invariant to selection and perturbation biases. This provides a unified perspective for understanding misalignment. Based on this, we further offer actionable insights into how misalignment should inform the design of real-world ML systems. We validate our theoretical findings through extensive empirical studies on both synthetic data and real image-text datasets, shedding light on the nuanced impact of misalignment on multimodal representation learning.

1 Introduction

Modern multimodal learning has achieved remarkable success by jointly modeling information from heterogeneous sources such as vision, language, and audio. In particular, multimodal contrastive learning (MMCL) on paired data has emerged as a dominant strategy for aligning modalities (Radford et al., 2021; Jia et al., 2021; Wu et al., 2022)—notably exemplified by vision-language models like CLIP, which learns a joint embedding space by maximizing the similarity of real image-text pairs while minimizing that of incorrect pairs (Radford et al., 2021). However, one fundamental assumption in multimodal learning is that the training pairs are perfectly aligned across modalities (Xia et al., 2023; Liu et al., 2024c). This assumption, though convenient, is often violated in real-world scenarios, where multimodal data is inherently noisy or imprecisely paired (Miech et al., 2019; Nakada et al., 2023), which we refer to as *misalignment*. For example, in a large-scale video-text dataset, over 50% of the purportedly aligned clip-caption pairs were found to be misaligned (Miech et al., 2019). Such misalignment, where the supposed counterparts (e.g., an image and its text) fail to correspond meaningfully, presents an unexpected and underexplored challenge in multimodal representation learning.

The misalignment discussed above has led to two seemingly opposing viewpoints. On one hand, misalignment is viewed as a form of disruption that should be mitigated Liu et al. (2024a); Sun et al. (2024); Wu et al. (2022); Chen et al. (2024); Tao et al. (2024); Kim et al. (2024); Xuan et al. (2022); Chartsias et al. (2020); Lv et al. (2021). For example, misalignment between modalities can result in "hallucination" in multimodal models

[†] These authors contributed equally. Correspondence to `yichao.cai@adelaide.edu.au`.

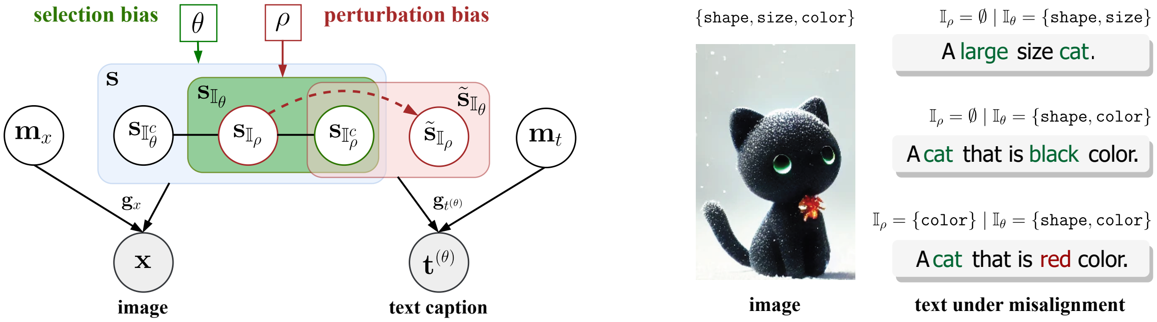

selection bias
θ
ρ
perturbation bias
$\mathbf{s}$
$\mathbf{s}_{\mathbb{I}_\theta}$
$\tilde{\mathbf{s}}_{\mathbb{I}_\theta}$
$\mathbf{m}_x$
$\mathbf{s}_{\mathbb{I}_\theta^c}$
$\mathbf{s}_{\mathbb{I}_\rho}$
$\mathbf{s}_{\mathbb{I}_\rho^c}$
$\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$
$\mathbf{m}_t$
$\mathbf{g}_x$
$\mathbf{g}_{t^{(\theta)}}$
$\mathbf{x}$
$\mathbf{t}^{(\theta)}$
image
text caption
{shape, size, color}
$\mathbb{I}_\rho = \emptyset \mid \mathbb{I}_\theta = \{$shape, size$\}$
A large size cat.
$\mathbb{I}_\rho = \emptyset \mid \mathbb{I}_\theta = \{$shape, color$\}$
A cat that is black color.
$\mathbb{I}_\rho = \{$color$\} \mid \mathbb{I}_\theta = \{$shape, color$\}$
A cat that is red color.
image
text under misalignment

Figure 1: **Illustration of the proposed latent variable model (left), with misalignment across modalities modeled via selection and perturbation bias.** Image $\mathbf{x}$ is generated from semantic variables $\mathbf{s}$ and image-specific variables $\mathbf{m}_x$ via the generative process $\mathbf{g}_x$. The corresponding text $\mathbf{t}^{(\theta)}$ is generated by $\mathbf{g}_{t^{(\theta)}}$ which acts on a biased subset of semantic variables $\tilde{\mathbf{s}}_{\mathbb{I}_\theta} = (\mathbf{s}_{\mathbb{I}_\rho^c}, \tilde{\mathbf{s}}_{\mathbb{I}_\rho})$, influenced by selection bias θ and perturbation bias ρ, along with text-specific variables $\mathbf{m}_t$. Selection bias omits $\mathbf{s}_{\mathbb{I}_\theta^c}$, while perturbation bias marks a subset $\mathbb{I}_\rho \subseteq \mathbb{I}_\theta$ whose components may be randomly replaced to form $\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$. Example image-text pairs (right) illustrate the misalignment induced by these two biases.

Sun et al. (2024); Leng et al. (2024); Xie et al. (2024). It also provides weak, noisy, and even misleading supervision for multimodal pre-training Wu et al. (2022); Zhang et al. (2024b). On the other hand, an opposing viewpoint suggests that multimodal representations may actually benefit from misalignment (Wu et al., 2023; Nakada et al., 2023; Cai et al., 2025; Kim et al., 2023). For instance, fine-tuning the representations learned by CLIP through random text augmentation—which deliberately introduces misalignment in style-related information—can lead to more robust representations for zero-shot learning, few-shot learning, and even adversarial attacks Cai et al. (2025). This contrast raises a crucial question:

How can we theoretically reconcile these two opposing views on misalignment,
and, more importantly, determine which should guide practical applications?

In light of this, we offer a theoretical perspective that not only facilitates the understanding of misalignment but also provides insights into real-world applications. Specifically, we formulate the problem via a latent variable model (LVM), which captures the underlying generative process of image-text data with misalignment, as shown in Figure 1. In it, the latent space consists of shared semantic variables representing factors common to both modalities (e.g., object shapes and colors), along with modality-specific subspaces that capture unique variations in images and text.

To model misalignment, we introduce two mechanisms: selection bias and perturbation bias. Both act on the shared semantic information but differ in their effects. Selection bias determines which semantic information is preserved in the text. For example, when describing an object, the text might preserve information about its color ("black") but omit details about its texture or shape. On the other hand, perturbation bias introduces errors, such as changing "black" (correct color) to "red" (incorrect color). To make the proposed LVM adaptable to a wide range of real-world scenarios, we allow for an arbitrary causal structure among the latent semantic variables, providing flexibility in multimodal contexts. Finally, given their differences, we model the modalities with separate generative processes.

Building on the proposed LVM, we present a theoretical identifiability analysis within the MMCL framework. We show, under mild assumptions, that the subset of semantic variables unaffected by selection and perturbation biases remain block-identifiable (Defn. 4.1)—that is, only the unaffected subset of semantic variables admits a nonlinear and invertible mapping to the representations learned by MMCL in the proposed LVM. In contrast, the remaining semantic variables that are affected by misalignment are inherently excluded from the learned representations, regardless of the latent causal structure among all semantic variables. This result provides a unified perspective on the seemingly opposing views discussed above. While misalignment can be problematic in tasks that rely on fully preserving semantic information to maximize downstream utility, it may paradoxically become beneficial in scenarios where robustness to distribution shifts is desired. In such cases, misalignment acts as a natural regularizer, implicitly guiding models to focus on stable, invariant factors shared across modalities.

We summarize our key contributions as follows, and a detailed discussion of related work is provided in § 6.

(i) We propose a latent variable model for multimodal data generation that explicitly captures misalignment through two mechanisms: selection bias and perturbation bias (§ 3).

(ii) We establish a general identifiability result, showing that MMCL recovers the subset of semantic variables unaffected by these biases, independent of the underlying latent causal structure (§ 4.1).

(iii) We extend this result to two practical scenarios—tasks requiring common representations and those targeting invariant representations—offering actionable insights into how misalignment should inform real-world applications (§ 4.2).

(iv) We empirically validate our theoretical findings through extensive experiments on both real-world and synthetic image-text datasets, across diverse misalignment settings, including those with structured latent dependencies (§ 5).

2 Preliminaries: Multimodal Contrastive Learning

Multimodal contrastive learning (MMCL) (Radford et al., 2021; Jia et al., 2021; Zhang et al., 2022) aims to learn joint representations by aligning paired samples from different modalities, i.e., $\mathbf{t} \in \mathcal{T}$ for text and $\mathbf{x} \in \mathcal{X}$ for images, while pushing apart unpaired (negative) samples. In practice, MMCL typically employs two modality-specific encoders, i.e., $\mathbf{f}_t(\mathbf{t})$ for text and $\mathbf{f}_x(\mathbf{x})$ for images which project observed paired data into a shared representation space. The learning objective is generally formulated as the following contrastive loss:

$$\mathcal{L}_{\text{MMCL}}(\mathbf{f}_x, \mathbf{f}_t) = -\frac{1}{2K}\left[\sum_{i=1}^{K} \log \frac{e^{\langle \mathbf{f}_x(\mathbf{x}_i), \mathbf{f}_t(\mathbf{t}_i)\rangle/\tau}}{\sum_{j=1}^{K} e^{\langle \mathbf{f}_x(\mathbf{x}_i), \mathbf{f}_t(\mathbf{t}_j)\rangle/\tau}} + \sum_{i=1}^{K} \log \frac{e^{\langle \mathbf{f}_x(\mathbf{x}_i), \mathbf{f}_t(\mathbf{t}_i)\rangle/\tau}}{\sum_{j=1}^{K} e^{\langle \mathbf{f}_x(\mathbf{x}_j), \mathbf{f}_t(\mathbf{t}_i)\rangle/\tau}}\right], \tag{1}$$

where $\{\mathbf{x}_i, \mathbf{t}_i\}_{i=1}^{K}$ are sampled paired data, K denotes the number of training pairs, τ is a temperature hyperparameter controlling the sharpness of the similarity distribution, and $\langle \cdot, \cdot \rangle$ represents a similarity measure. Asymptotically, when K approaches infinity, and with $\tau = 1$ and the similarity function defined as the negative squared Euclidean distance, the objective in Eq. (1) reduces to (Daunhawer et al., 2022):

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x, \mathbf{f}_t) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{t})\sim p_{\mathbf{x},\mathbf{t}}} \left[\|\mathbf{f}_x(\mathbf{x}) - \mathbf{f}_t(\mathbf{t})\|_2\right] - \frac{1}{2}\Big(H\big(\mathbf{f}_x(\mathbf{x})\big) + H\big(\mathbf{f}_t(\mathbf{t})\big)\Big), \tag{2}$$

where $H(\cdot)$ denotes differential entropy (Wang and Isola, 2020; Von Kügelgen et al., 2021). One of the main advantages of the asymptotic objective in Eq. (2) is its suitability for theoretical analysis (Lyu et al., 2022; Daunhawer et al., 2022; Yao et al., 2023). At a high level, Eq. (2) naturally decomposes into two intuitive terms: the first term encourages minimizing the distance between paired samples, while the second term promotes maximizing the entropy of learned representations. Following prior works (Lyu et al., 2022; Daunhawer et al., 2022; Yao et al., 2023), we also adopt this objective in our theoretical analysis.

However, a key distinction from prior studies lies in our focus on the effect of *misalignment*—whereas existing works typically assume perfect alignment between paired data. To capture the impact of misalignment, we introduce a novel latent variable model, as discussed in § 3, which leads to a fundamentally different problem setting. As a result of this distinct problem context, our theoretical results also differ substantially from prior work, offering new insights into how misalignment shapes the learned representations in multimodal contrastive learning, as shown in § 4.

3 Problem Formulation via a Generative Perspective

In this section, we introduce a novel latent variable model (LVM) to formally characterize misalignment (§ 3.1). The model incorporates two key mechanisms—selection bias and perturbation bias—to explicitly capture distinct sources of misalignment. Building on this model, we present technical assumptions underlying image-text pairs under misalignment (§ 3.2).

3.1 A Latent Variable Model Characterizing Misalignment

Figure 1 illustrates the proposed LVM. In the following, we provide a detailed explanation of the model from three aspects: the latent space, image generation, and text generation.

Latent Space. We partition the entire latent space $\mathcal{Z}$ into three simply connected, open subspaces, i.e., $\mathcal{Z} = \mathcal{S} \times \mathcal{M}_x \times \mathcal{M}_t$, where each defines the support of a distinct group of latent variables. We denote the latent variables in $\mathcal{Z}$ as $\mathbf{z} = (\mathbf{s}, \mathbf{m}_x, \mathbf{m}_t)$, where $\mathbf{s}$, $\mathbf{m}_x$, and $\mathbf{m}_t$ lie in $\mathcal{S}$, $\mathcal{M}_x$, and $\mathcal{M}_t$, respectively. Below, we describe the characteristics of the latent variables $\mathbf{s}$, $\mathbf{m}_x$, and $\mathbf{m}_t$:

- *Semantic variables* $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^{n_s}$: Latent variables capturing the semantic content of the data, i.e., information that is interpretable or describable through human knowledge (e.g., object shape, color). We denote the index set of semantic variables as $\mathbb{I}_\mathbf{s} := \{1, \dots, n_s\}$ for future reference.

- *Image-specific variables* $\mathbf{m}_x \in \mathcal{M}_x$ [1]: Latent variables capturing image-specific, non-semantic factors (e.g., camera noise or background textures) that are independent of semantic variables $\mathbf{s}$.

- *Text-specific variables* $\mathbf{m}_t \in \mathcal{M}_t$: Latent variables capturing text-specific, non-semantic factors (e.g., grammar) that are independent of both semantic variables $\mathbf{s}$ and image-specific variables $\mathbf{m}_x$.

A key advantage of the proposed LVM is its flexibility and applicability to real-world scenarios. To this end, we depart from prior works that impose somewhat restrictive assumptions on the latent structure. Specifically, unlike approaches that enforce certain fixed graphical structures (e.g., assuming content causally determines style) (Von Kügelgen et al., 2021; Daunhawer et al., 2022), or methods based on nonlinear ICA that assume complete independence among latent variables (Khemakhem et al., 2020; Sorrenson et al., 2020), we allow for arbitrary dependency structures among semantic variables $\mathbf{s}$, since the true latent graph structure is often unknown in practice.

Image Generation. Images $\mathbf{x} \in \mathcal{X}$ are generated from latent variables $\mathbf{z}_x = (\mathbf{s}, \mathbf{m}_x)$ via a diffeomorphism (i.e., a bijection with smooth inverse function) $\mathbf{g}_x : \mathcal{S} \times \mathcal{M}_x \to \mathcal{X}$:

$$\mathbf{s} \sim p_\mathbf{s}, \quad \mathbf{m}_x \sim p_{\mathbf{m}_x}, \quad \mathbf{z}_x = (\mathbf{s}, \mathbf{m}_x), \quad \mathbf{x} = \mathbf{g}_x(\mathbf{z}_x). \tag{3}$$

Here, $p_\mathbf{s}$ and $p_{\mathbf{m}_x}$ are prior distributions over subspaces $\mathcal{S}$ and $\mathcal{M}_x$, respectively. $\mathcal{X}$ is a smooth manifold that defines the image observation space. This generative process formalizes that images fully encapsulate semantics, reflecting the fact that they are both informative and semantically rich.

Text Generation. Unlike other multimodal data (e.g., camera-LiDAR (Liang et al., 2022)) acquired via sensors, given an image, its corresponding text inherently exhibits flexibility in semantic richness and may also include distortions introduced by humans or captioning models (Li et al., 2022; Xu et al., 2024), leading to misalignment. Here, we formalize such misalignment by introducing two types of biases: selection bias and perturbation bias. Accordingly, before formulating the text generation process, we first provide definitions of these two types of biases.

Definition 3.1 (Selection Bias θ)**.** Let $\mathcal{P}_+(\mathbb{I}_\mathbf{s})$ denote the set of all non-empty subsets[2] of the index set $\mathbb{I}_\mathbf{s}$, defined as $\mathcal{P}_+(\mathbb{I}_\mathbf{s}) := \mathcal{P}(\mathbb{I}_\mathbf{s}) \setminus \{\emptyset\}$, where $\mathcal{P}(\cdot)$ denotes the power set. The *selection bias* θ is defined as an integer index in the range $\theta \in \{1, \dots, 2^{n_s} - 1\}$, corresponding to a specific non-empty semantic subset $\mathbb{I}_\theta \in \mathcal{P}_+(\mathbb{I}_\mathbf{s})$. The complement $\mathbb{I}_\theta^c = \mathbb{I}_\mathbf{s} \setminus \mathbb{I}_\theta$ denotes the omitted semantic subset.

Note that each θ uniquely determines a non-empty semantic subset $\mathbb{I}_\theta \in \mathcal{P}_+(\mathbb{I}_\mathbf{s})$, which defines the semantic information to be expressed in the generated text. It also specifies a text generation mapping $\mathbf{g}_{t^{(\theta)}} : \mathcal{S}_{\mathbb{I}_\theta} \times \mathcal{M}_t \to \mathcal{T}^{(\theta)}$, selected from a class of diffeomorphisms $\mathcal{G}_t$. Here, $\mathcal{T}^{(\theta)}$ denotes a smooth manifold that contains the observed text $\mathbf{t}^{(\theta)}$ under selection bias θ.

Definition 3.2 (Perturbation Bias ρ)**.** Let $\mathcal{P}_{\text{proper}}(\mathbb{I}_\theta) := \mathcal{P}(\mathbb{I}_\theta) \setminus \{\mathbb{I}_\theta\}$ denote the set of all proper subsets[3] of the selected index subset $\mathbb{I}_\theta$. The *perturbation bias* ρ is defined as an integer index in the range $\rho \in \{1, \dots, 2^{|\mathbb{I}_\theta|} - 1\}$, corresponding to a unique subset $\mathbb{I}_\rho \in \mathcal{P}_{\text{proper}}(\mathbb{I}_\theta)$ subject to perturbation. The complement $\mathbb{I}_\rho^c := \mathbb{I}_\theta \setminus \mathbb{I}_\rho$ denotes the semantic components that remain fixed.

[1] For simplicity of notation, we omit the dimensions of certain variables throughout this work.

[2] Without loss of generality, we fix a graded lexicographic order over $\mathcal{P}_+(\mathbb{I}_\mathbf{s})$, induced by the order in $\mathbb{I}_\mathbf{s}$.

[3] Again, we fix a graded lexicographic order over the proper subsets in $\mathcal{P}_{\text{proper}}(\mathbb{I}_\theta)$ throughout the paper.

Example 3.1. Let the full semantic index set be $\mathbb{I}_{\mathbf{s}}$ $=\{\texttt{shape}, \texttt{size}, \texttt{color}\}$, so that the set of non-empty semantic subsets is $\mathcal{P}_{+}(\mathbb{I}_{\mathbf{s}}) = \{\{\texttt{shape}\}, \{\texttt{size}\}, \ldots, \{\texttt{shape}, \texttt{size}, \texttt{color}\}\}$. Then, a selection bias $\theta = 5$ corresponds to the fifth subset, i.e., $\mathbb{I}_\theta = \{\texttt{shape}, \texttt{color}\}$, with the omitted semantics given by $\mathbb{I}_\theta^c = \{\texttt{size}\}$. The corresponding text-generation mapping $\mathbf{g}_{t^{(\theta)}}$ uses only the selected semantic variables in $\mathbb{I}_\theta$ to generate text $\mathbf{t}^{(\theta)}$. The set of proper subsets of $\mathbb{I}_\theta$ is $\mathcal{P}_{\text{proper}}(\mathbb{I}_\theta) = \{\emptyset, \{\texttt{shape}\}, \{\texttt{color}\}\}$. Then, a perturbation bias $\rho = 3$ corresponds to the subset $\mathbb{I}_\rho = \{\texttt{color}\}$, and its complement within $\mathbb{I}_\theta$ is $\mathbb{I}_\rho^c = \{\texttt{shape}\}$. Under the combined biases $\theta = 5$ and $\rho = 3$, only shape is unbiasedly preserved, while color is subject to perturbation. A resulting text might be *"A cat that is red color"*, even though the image shows a large-sized black cat.

Building upon the previously defined selection bias θ and perturbation bias ρ, we now formalize the text generation process, explicitly capturing the misalignment induced by these two biases. Consider an image $\mathbf{x}$ generated by Eq. (3), with associated semantic variables $\mathbf{s} = (\mathbf{s}_{\mathbb{I}_\theta}, \mathbf{s}_{\mathbb{I}_\theta^c})$, where the index set $\mathbb{I}_\theta$ is determined by θ. For the corresponding text $\mathbf{t}^{(\theta)}$, we define the latent variables as $\mathbf{z}_{t^{(\theta)}} = (\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t)$, where $\tilde{\mathbf{s}}_{\mathbb{I}_\theta}$ represents the perturbed semantic variables under perturbation bias ρ, and $\mathbf{m}_t$ denotes the text-specific latent variables. The text generation process is then formalized as:

$$\tilde{\mathbf{s}}_{\mathbb{I}_\theta} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}, \quad \mathbf{m}_t \sim p_{\mathbf{m}_t}, \quad \mathbf{z}_{t^{(\theta)}} = (\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t), \quad \mathbf{t}^{(\theta)} = \mathbf{g}_{t^{(\theta)}}(\mathbf{z}_{t^{(\theta)}}), \tag{4}$$

where $p_{\mathbf{m}_t}$ denotes the prior distribution over the latent subspace $\mathcal{M}_t$, and $\mathbf{g}_{t^{(\theta)}}$ is the diffeomorphic mapping specified by the selection bias θ. This formulation explicitly captures how misalignment in text generation arises through perturbations within selected semantic dimensions.

3.2 Model Assumptions for Theoretical Analysis

We now present the technical assumptions underlying our theoretical analysis, based on the proposed LVM:

Assumption 3.1 (Latent Variables with Continuous Positive Densities)**.** *The latent variables* $\mathbf{s}$*,* $\mathbf{m}_x$*, and* $\mathbf{m}_t$ *are continuous and admit strictly positive densities, i.e.,* $p_{\mathbf{s}} > 0$*,* $p_{\mathbf{m}_x} > 0$*, and* $p_{\mathbf{m}_t} > 0$*, almost everywhere (a.e.) on their respective supports* $\mathcal{S}$*,* $\mathcal{M}_x$*, and* $\mathcal{M}_t$*.*

Assumption 3.2 (Random Perturbations)**.** *Given a selection bias* θ *and a perturbation bias* ρ*, consider an image-text pair* $(\mathbf{x}, \mathbf{t}^{(\theta)})$ *generated by Eq.* (3) *and Eq.* (4)*, respectively. The conditional distribution* $p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}$ *is defined via a randomly sampled perturbation subset* $A \subseteq \mathbb{I}_\rho$*, such that:*

$$A \sim p_A, \quad p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}(\tilde{\mathbf{s}}_{\mathbb{I}_\theta} \mid \mathbf{s}, A) = \delta\big(\tilde{\mathbf{s}}_{\mathbb{I}_\theta \setminus A} - \mathbf{s}_{\mathbb{I}_\theta \setminus A}\big) \cdot p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A). \tag{5}$$

Here: **(i)** *A is the random subset of semantic indices to be perturbed, with p_A defined over $\mathcal{P}(\mathbb{I}_\rho)$. For every $l \in \mathbb{I}_\rho$, there exists at least one subset $A \subseteq \mathbb{I}_\rho$ such that $l \in A$ and $p_A(A) > 0$;* **(ii)** *$\delta(\cdot)$ denotes the Dirac delta function, enforcing that variables outside A remain unchanged;* **(iii)** *$p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a smooth, strictly positive conditional density over $\mathcal{S}_A \times \mathcal{S}_A$, where $\mathcal{S}_A$ is the domain of $\mathbf{s}_A$, and for each $\mathbf{s}_A$, the support of $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ includes a non-empty open subset $\mathcal{O}_A \subseteq \mathcal{S}_A$.*

Interpretation 3.1. Note that Eq. (5) essentially implies that, in each $(\mathbf{x}, \mathbf{t}^{(\theta)})$, only a subset of semantic variables A undergoes perturbations, regardless of the underlying causal structure among latent semantic variables. The rationale is that latent semantic variables can only be modified indirectly by altering observations, rather than through direct intervention, unless the latent causal structure is fully identified. In the text modality, it occurs through the misassignment of certain content words to specific image semantics during the captioning process. Unlike a direct intervention, this misalignment does not propagate to descendant semantic variables, as illustrated in Figure 2.

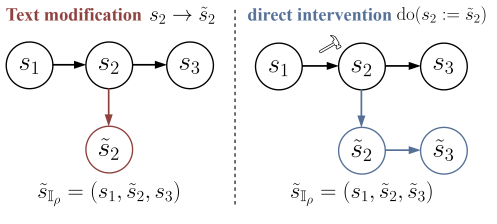

Text modification $s_2 \rightarrow \tilde{s}_2$
direct intervention $\mathrm{do}(s_2 := \tilde{s}_2)$
s_1
s_2
s_3
$\tilde{s}_2$
$\tilde{s}_{\mathbb{I}_\rho} = (s_1, \tilde{s}_2, s_3)$
s_1
s_2
s_3
$\tilde{s}_2$
$\tilde{s}_3$
$\tilde{s}_{\mathbb{I}_\rho} = (s_1, \tilde{s}_2, \tilde{s}_3)$

Figure 2: **Text modifications** affect only the semantics where content words are changed, while direct interventions act on latent semantic variables thereby propagating structural changes.

4 Theory: Identifiability Results of Latent Semantic Variables

In this section, we theoretically analyze how misalignment impacts the identifiability of latent semantic variables in the proposed LVM, within the MMCL framework (§ 4.1). Based on these results, we further provide practical insights into how misalignment should be addressed in real-world applications (§ 4.2). Detailed proofs of the theoretical results are provided in App. B.

To begin, we restate the definition of *block-identifiability* (Von Kügelgen et al., 2021) in the context of our problem setting:

Definition 4.1 (Block-Identifiability)**.** A subset of latent semantic variables $\mathbf{s}_{\mathbb{I}_{\mathrm{id}}} \in \mathcal{S}_{\mathbb{I}_{\mathrm{id}}}$, with $\mathbb{I}_{\mathrm{id}} \subseteq \mathbb{I}_{\mathbf{s}}$, is said to be *block-identified* by functions $\mathbf{f}_x : \mathcal{X} \to \mathbb{R}^{|\mathbb{I}_{\mathrm{id}}|}$ and $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to \mathbb{R}^{|\mathbb{I}_{\mathrm{id}}|}$ if the learned representations $\hat{\mathbf{z}}_x \in \mathbb{R}^{|\mathbb{I}_{\mathrm{id}}|}$ and $\hat{\mathbf{z}}_t \in \mathbb{R}^{|\mathbb{I}_{\mathrm{id}}|}$ retain *all and only* the information contained in $\mathbf{s}_{\mathbb{I}_{\mathrm{id}}}$. Formally, there exist invertible mappings $\mathbf{h}_x, \mathbf{h}_t : \mathcal{S}_{\mathbb{I}_{\mathrm{id}}} \to \mathbb{R}^{|\mathbb{I}_{\mathrm{id}}|}$ s.t. $\hat{\mathbf{z}}_x = \mathbf{h}_x(\mathbf{s}_{\mathbb{I}_{\mathrm{id}}})$ and $\hat{\mathbf{z}}_t = \mathbf{h}_t(\mathbf{s}_{\mathbb{I}_{\mathrm{id}}})$.

4.1 Identifiability Result of Latent Semantic Variables under Misalignment

Building on the definition above and the parameterization and assumptions outlined in § 3 for the proposed LVM, we provide the following identifiability result:

Theorem 4.1 (Identifiability of Latent Semantic Variables)**.** *Let* $(\mathbf{x}, \mathbf{t}^{(\theta)})$ *be image-text pairs drawn from the data-generating process described in § 3, where* $\mathbf{x}$ *is generated according to Eq.* (3) *and* $\mathbf{t}_\theta$ *is generated by Eq.* (4)*. Suppose that Asms. 3.1 and 3.2 hold. Denote by* $\mathbf{s}_{\mathbb{I}_\rho^c}$ *the subset of semantic variables that preserved without bias in the text, and define its dimension as* $n = |\mathbb{I}_\theta| - |\mathbb{I}_\rho|$*. Let* $\mathbf{f}_x : \mathcal{X} \to (0,1)^n$ *and* $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^n$ *be sufficiently flexible, smooth functions. Then, Minimizing the loss* $\mathcal{L}_{\mathit{SymAlignMaxEnt}}$ *in Eq.* (2) *over samples* $(\mathbf{x}, \mathbf{t}_\theta)$ *guarantees that* $\mathbf{f}_x$ *and* $\mathbf{f}_t$ *block-identify the semantic variables* $\mathbf{s}_{\mathbb{I}_\rho^c}$ *in the sense of Defn. 4.1.*

Discussion 4.1. Thm. 4.1 formally establishes that, in the presence of misalignment, the unbiased semantic variables $\mathbf{s}_{\mathbb{I}_\rho^c}$ that are shared across modalities can be effectively recovered, up to a block-wise indeterminacy, by minimizing the MMCL objective. In contrast, components that are misaligned, specifically $\mathbf{s}_{\mathbb{I}_\rho}$ and $\mathbf{s}_{\mathbb{I}_\theta^c}$, are entirely excluded from the learned representations. We emphasize that this result holds regardless of any underlying latent graph structure among semantic variables. At its core, this result highlights the model's capacity to focus exclusively on the aligned semantic aspects of the data. Furthermore, modality-specific, non-semantic factors, i.e., $\mathbf{m}_x$ and $\mathbf{m}_t$, are consistently discarded throughout the learning process. This further underscores the model's ability to extract meaningful, cross-modal semantic information while filtering out semantically uninformative or noise-induced components.

4.2 Insights into Misalignment for Practice

The above result is general and not limited to any specific problem context. We now consider two real-world application scenarios: **[i]** pretraining with large-scale data and **[ii]** invariant representation learning. The former aims to capture comprehensive semantic information to support a wide range of downstream tasks, while the latter focuses on learning robust representations for out-of-distribution (OOD) generalization. In what follows, we present corollaries and insights for both scenarios.

Corollary 4.1 (Identifiability of Full Latent Semantic Variables)**.** *Let the selection bias be* $\theta = 2^{n_s} - 1$ *and the perturbation bias be* $\rho = 1$*, such that the full set of semantic variables* $\mathbb{I}_{\mathbf{s}}$ *is selected, and the perturbable semantic subset is trivial, i.e.,* $\mathbb{I}_\rho = \emptyset$*. Then, all semantic variables* $\mathbf{s}$ *are block-identified via smooth functions* $\mathbf{f}_x : \mathcal{X} \to (0,1)^{n_s}$ *and* $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^{n_s}$*, when minimizing* $\mathcal{L}_{\mathit{SymAlignMaxEnt}}$*.*

Insight 4.1 (MMCL Pretraining on Large-Scale Data)**.** Large-scale multimodal datasets (e.g., COCO (Lin et al., 2014), Conceptual Captions (Sharma et al., 2018), LAION-5B (Schuhmann et al., 2022)) often exhibit varying caption quality. Our analysis indicates that omitted or perturbed semantics are irretrievably lost in the learned representations, regardless of dataset size, although scale may mitigate sporadic misalignment by averaging its effects. Preserving a breadth of relevant semantic details is therefore crucial when pretraining foundation models, whose primary goal is to support diverse downstream tasks. As noted in Cor. 4.1, achieving this requires detailed and consistent annotation of image semantics. Consequently, improved caption control (Li et al., 2022; Fang et al., 2015; Ding et al., 2023) is essential to avoid blind spots in semantic coverage.

Corollary 4.2 (Identifiability of Invariant Semantic Variables)**.** *Consider an OOD setting in which a subset of semantic variables,* $\mathbb{I}_{inv} \subset \mathbb{I}_{\mathbf{s}}$*, remains invariant between training and testing environments, while the remaining semantic variables,* $\mathbb{I}_{var} = \mathbb{I}_{\mathbf{s}} \setminus \mathbb{I}_{inv}$*, undergo distribution shifts. If the union of omitted and perturbable semantic variables under selection bias* θ *and perturbation bias* ρ *coincides with the environment-sensitive subset, i.e.,* $\mathbb{I}_{var} = \mathbb{I}_\theta^c \cup \mathbb{I}_\rho$*, then the invariant semantic variables* $\mathbf{s}_{\mathbb{I}_{inv}}$ *are block-identified via smooth functions* $\mathbf{f}_x : \mathcal{X} \to (0,1)^{|\mathbb{I}_{inv}|}$ *and* $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^{|\mathbb{I}_{inv}|}$*, by minimizing* $\mathcal{L}_{\mathit{SymAlignMaxEnt}}$*.*

Insight 4.2 (Invariant Representation Learning)**.** In tasks requiring robust OOD performance (e.g., domain generalization) (Rojas-Carulla et al., 2018; Arjovsky et al., 2019), semantic variables that are vulnerable to distribution shifts can undermine generalization. As noted in Cor. 4.2, misalignment may, counterintuitively, enhance robustness by selectively omitting or perturbing these vulnerable variables. This suggests that MMCL may offer a novel perspective on invariant representation learning (Peters et al., 2016; Krueger et al., 2021; Eastwood et al., 2022), as auditing and curating text is more precise and interpretable, since language is distilled from human knowledge.

5 Experiments

We conduct extensive experiments under diverse misalignment settings to validate our theoretical results, including numerical simulations (§ 5.1), a real-world image-text dataset with independent semantic variables (§ 5.2), and a synthetic dataset with dependent semantic variables. (§ 5.3).

5.1 Numerical Simulation

Experimental Setup. We synthesize numerical data following the generative process described in § 3. Specifically, we sample modality-specific variables $\mathbf{m}_x \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_x})$ and $\mathbf{m}_t \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_t})$, each of dimension 5, along with full semantic variables $\mathbf{s} \sim \mathcal{N}(0, \Sigma_{\mathbf{s}})$, of dimension 10. Potential causal dependencies among these variables are encoded in their covariance matrices $\Sigma_{(\cdot)}$. To simulate various misalignment scenarios, we progressively increase the strength of selection and perturbation biases. For ***selection bias***, we incrementally define selected subsets $\mathbb{I}_\theta$ as $\{1\}, \ldots, [10]$.[4] For ***perturbation bias***, we similarly use increasing subsets of $\mathbb{I}_\rho$, ranging from the empty set to $[9]$, applying additive Gaussian noise with a probability 0.75 for each semantic dimension $i \in \mathbb{I}_\rho$. These biases are applied separately: when analyzing selection bias, we set $\mathbb{I}_\rho = \emptyset$; conversely, when analyzing perturbation bias, we fix $\mathbb{I}_\theta = [9]$. Generation functions $\mathbf{g}_x$ and $\mathbf{g}_{t_\theta}$ are instantiated as randomly initialized invertible MLPs. Two modality-specific MLP encoders are trained for 100,000 steps using the $\mathcal{L}_{\text{SymAlignMaxEnt}}$ loss defined in Eq. (2), setting the representation dimension equal to the unbiased semantic dimensions. Further details on the experimental setup are provided in App. C.1.

We conduct two main experiments. **(i)** ***Identification of semantics***: We predict each dimension of the true semantic variables from the learned representations using a lightweight MLP, reporting the predictive R^2 score on holdout data. Results are averaged over three random seeds for each setting. **(ii)** ***Downstream performance***: We evaluate the pretrained representations obtained under various bias conditions on downstream tasks. Specifically, we construct four regression task labels y_1, y_2, y_3, and y_4 by applying complex nonlinear functions to subsets of the semantic variables: [3], [5], [7], and [9], respectively. For classification, we binarize y_2 to produce binary labels. To evaluate the OOD generalization performance of the trained classifiers, we introduce a distribution shift in the observations $\mathbf{x}$ by applying a heavy-tailed transformation to dimensions $\{9, 10\}$. Further details on the task design are provided in App. C.1.

Identification of Semantics. The results in Figure 3 show that, under the independent latent variable scenario, unbiased semantic variables are clearly block-identified ($R^2 \approx 1$), whereas misaligned semantics due to selection bias are effectively discarded ($R^2 \approx 0$). In the dependent latent variable scenario, some misaligned semantics become partially predictable, reflecting inherent mutual predictability among strongly dependent variables Von Kügelgen et al. (2021); Yao et al. (2023). Modality-specific variables are consistently omitted from the representations across all settings. Similar effects are observed under perturbation bias, as illustrated in Figure 7. Notably, although our theoretical results hold only up to invertible mappings, simple linear regression already achieves high R^2 scores, as demonstrated in Figure 8. These findings consolidate the identifiability results in Thm. 4.1. Additional analyses on misassigned encoding dimensions and combined bias effects, are provided in App. C.2.

Downstream Performance. As shown in Figure 4, retaining more semantic information during pretraining significantly enhances in-distribution regression performance, consistent with Cor. 4.1. Conversely, under distribution shift scenarios, accurately identifying invariant semantic variables is essential for robust out-of-distribution generalization. In this setting, introducing appropriate selection or perturbation biases effectively removes variables sensitive to distribution shift, supporting the result in Cor. 4.2. Additional results on the effects of perturbation bias are provided in App. C.3.

[4] For any positive integer $n > 1$, we use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

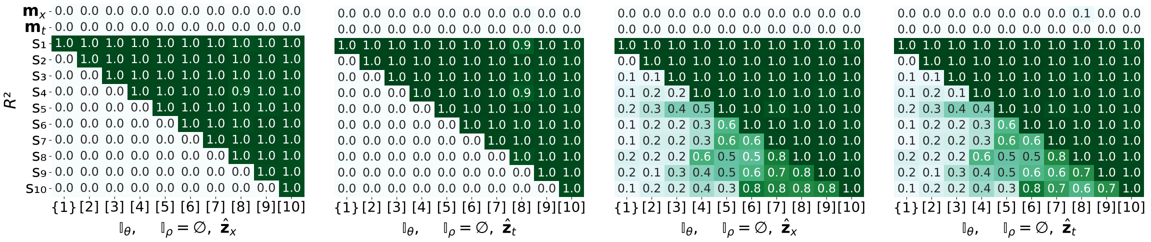


Figure 3: **Mean R^2 scores under selection bias settings.** From left to right: predictions based on $\hat{\mathbf{z}}_x$ and $\hat{\mathbf{z}}_t$ with independent latent semantics, followed by those with dependent latent semantics.

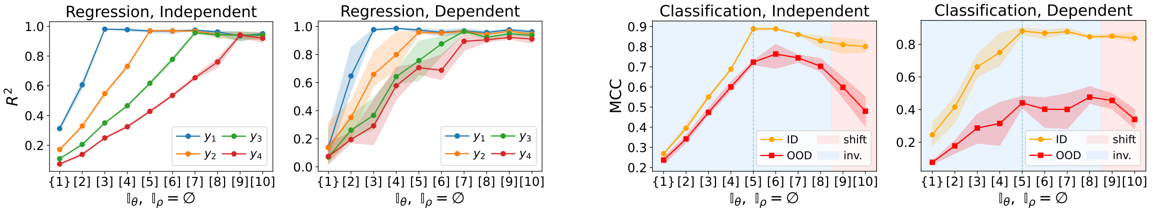


Figure 4: **Downstream performance of pretrained $\hat{\mathbf{z}}_x$ under selection bias.** Left: in-distribution (ID) regression performance. Right: ID classification and out-of-distribution (OOD) generalization.

5.2 MPI3d-Complex Dataset

Experimental Setup. The *MPI3D-Complex* dataset (Gondal et al., 2019) consists real-world images captured in a controlled environment, with mutually *independent, discrete* latent factors: object color (`color`; 4 values), shape (`shape`; 4 values), size (`size`; 2 values), camera height (`cam.`; 3 values), background color (`back.`; 3 values), and object position along the horizontal (`hori.`; 40 values) and vertical (`vert.`; 40 values) axes. We designate the positional factors (`hori.`, `vert.`) as image-specific, while the remaining attributes are treated as semantic variables. Text observations are generated based on the ground-truth semantic variables, with variation introduced via different selection and perturbation bias settings, and augmented with template-based text generation to simulate text-specific factors.

To systematically study the effects of misalignments, we define incremental subsets for each setting. For ***selection bias*** ($\mathbb{I}_\theta$), we use the following configurations, all with no perturbation ($\mathbb{I}_\rho = \emptyset$): ①: {`color`}, ②: {`color`, `shape`}, ③: {`color`, `shape`, `size`}, ④: {`color`, `shape`, `size`, `cam.`}, and ⑤: {`color`, `shape`, `size`, `cam.`, `back.`}. Conversely, for ***perturbation bias*** ($\mathbb{I}_\rho$), we define the following configurations, all with a fixed full selection of semantic attributes ($\mathbb{I}_\theta = \{$`color`, `shape`, `size`, `cam.`, `back.`$\}$): ①: $\emptyset$, ②: {`back.`}, ... up to ⑤: {`shape`, `size`, `cam.`, `back.`}.

Training is performed using the multimodal contrastive loss $\mathcal{L}_{\text{MMCL}}$ defined in Eq. (1), following the formulation in (Daunhawer et al., 2022). Performance is evaluated using the average Matthews Correlation Coefficient (MCC) across three random seeds, based on the prediction of all image latent variables from the learned image and text representations using linear and nonlinear MLP decoders, respectively. Further implementation details—including dataset statistics, bias configurations, text generation procedures, encoder architectures, and training hyperparameters—are provided in App. D.1.

Results. Figure 5 presents the MCC results using nonlinear classifiers with learned representations. The findings demonstrate that, even with discrete latent variables, misaligned semantic variables across modalities, whether due to selection or perturbation biases, are systematically excluded from the representations (MCC $= 0$). In contrast, unbiased semantics are well recovered, with MCC scores predominantly approaching 1 and all values ≥ 0.8, reinforcing our theoretical findings. Further investigations into MCC using linear classifiers and ablation studies on encoder dimensionality are provided in App. D.2.

5.3 Causal3DIdent Dataset

Experimental Setup. We conduct our experiments using the *Causal3DIdent* dataset, a semi-synthetic benchmark widely used in the causal representation learning literature (Zimmermann et al., 2021; Von Kügelgen et al., 2021; Daunhawer et al., 2022; Yao et al., 2023). This dataset enables explicit enforcement of causal structures over latent variables. The generative process for images is governed by a set of 10 latent variables,

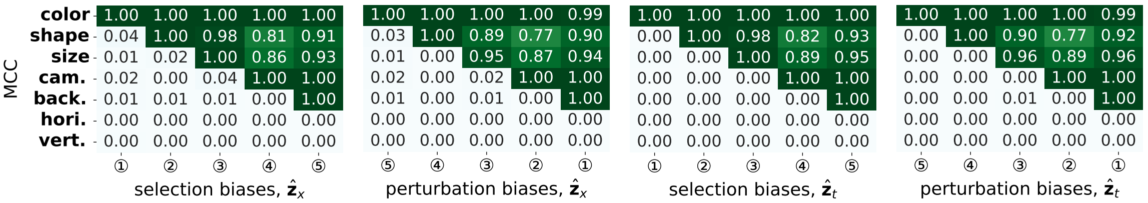


Figure 5: **Mean MCC scores under misalignment settings.** Left to right: Image features $\hat{\mathbf{z}}_x$ with selection and perturbation bias settings, text features $\hat{\mathbf{z}}_t$ under the same bias settings.

comprising 3 discrete factors—object shape (`shape`), and object positions along the horizontal (`x_pos`) and vertical (`y_pos`) axes—and 7 continuous factors: object color (`color`); spotlight position (`s_pos`) and color (`s_color`); background color (`b_color`); and object rotation angles (`alpha`, `beta`, `gamma`). We treat the rotation angles (`alpha`, `beta`, `gamma`) as image-specific variables, and the remaining factors as semantic variables, which follow a predefined causal structure illustrated in Figure 14.

For text latent semantic variables, we discretize `color`, `s_color` and `b_color`, while keeping `s_pos` continuous, and simulating partial information loss in spotlight position when mapping to text observations. For text-specific variables, we use five manually designed templates and generate text from latent variables according to each bias setting, adapting the text rendering process from (Daunhawer et al., 2022). We consider the following ***selection bias*** settings ($\mathbb{I}_\theta$): ① refers to $\{\texttt{shape}\}$; ②, $\{\texttt{shape, x\_pos}\}$; ③, $\{\texttt{shape}, \texttt{x\_pos}, \texttt{y\_pos}\}$; ④, $\{\texttt{shape}, \texttt{x\_pos}, \texttt{y\_pos}, \texttt{s\_pos}\}$; ⑤, $\{\texttt{shape}, \texttt{x\_pos}, \texttt{y\_pos}, \texttt{s\_pos}, \texttt{color}\}$; ⑥, $\{\texttt{shape}, \texttt{x\_pos}, \texttt{y\_pos}, \texttt{s\_pos}, \texttt{color}, \texttt{s\_color}\}$; and ⑦, $\{\texttt{shape}, \texttt{x\_pos}, \texttt{y\_pos}, \texttt{s\_pos}, \texttt{color}, \texttt{s\_color}, \texttt{b\_color}\}$, all with $\mathbb{I}_\rho = \emptyset$. For ***perturbation bias*** ($\mathbb{I}_\rho$), we consider the reversed setup: ① corresponds to $\emptyset$; ② to $\{\texttt{b\_color}\}$; ... up to ⑦ to $\{\texttt{x\_pos}, \texttt{y\_pos}, \texttt{s\_pos}, \texttt{color}, \texttt{s\_color}, \texttt{b\_color}\}$, all with full semantic selection.

We synthesize 80,000 samples for MMCL pretraining, 10,000 samples for training classifiers or regressors, and additional 10,000 samples for evaluating predictive performance. Evaluation metrics include R^2 for continuous latent variables and MCC for discrete latent variables. Further details on dataset specifications, encoder architectures, and training parameters, are provided in App. E.1.

Results. Figure 6 presents the prediction performance of a nonlinear MLP classifier or regressor trained on learned image representations. We observe that unbiased semantic variables, whether continuous (e.g., `s_pos`) or discrete (e.g., `shape`, `x_pos`, `y_pos`), are reliably captured across all settings, with predictive performance approaching perfect ($R^2 \approx 1$). For semantic variables that are continuous in the image modality but discretized in the text modality, prediction performance shows some degradation (e.g., `s_color` in selection setting ⑥ or perturbation setting ②), yet still achieves relatively high R^2 scores. Image-specific variables are consistently excluded from the learned representations, as indicated by $R^2 = 0$. Likewise, semantic variables omitted due to selection or perturbation bias are generally discarded. For instance, in selection setting ① or perturbation setting ⑦, only `shape` remains predictable. When factors such as `x_pos` or `y_pos` are included, other dependent semantic variables, such as `color`, become partially predictable (e.g., in selection setting ②). Similarly, identification of `s_pos` enhances predictability of `s_color` and `b_color`, reflecting the latent causal structure illustrated in Figure 14. Overall, the results on the Causal3DIdent dataset further support our theoretical findings. Analyses of the text representations are provided in App. E.2.

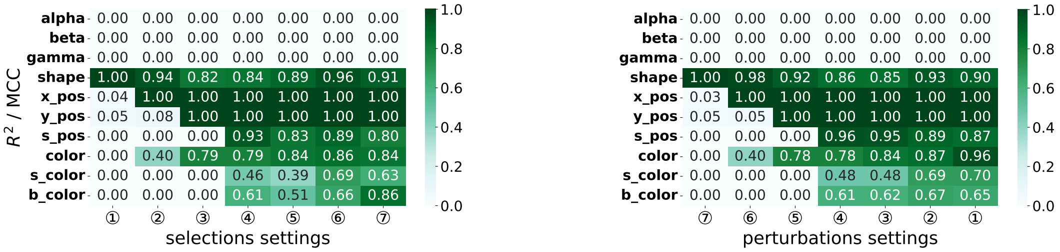


Figure 6: **Predicting semantic variables under misalignment using image features.** R^2 is reported for continuous factors and MCC for discrete factors. Left: selection bias. Right: perturbation bias.

6 Related Work

We now contextualize our work within four key lines of related research: theoretical analysis of contrastive learning, vision–language models, identifiability in latent variable models, and invariant representation learning.

Theoretical Multimodal (Multi-View) Contrastive Learning. Recent work has sought to formalize the theoretical foundations of multimodal and multi-view representation learning, particularly under contrastive objectives. Wang et al. Wang and Isola (2020) decompose the InfoNCE loss Oord et al. (2018) into an alignment term, which pulls positive pairs together, and a uniformity term, which encourages dispersion over a hypersphere—laying the groundwork for subsequent analysis. Zimmermann et al. Zimmermann et al. (2021) show that contrastive objectives can invert the data-generating process, while Liu et al. Liu et al. (2024c) extend this result to multimodal settings. However, these approaches often rely on strong assumptions about latent distributions or manifold structures, which limits their practical applicability. A complementary line of work, such as Von Kügelgen et al. (2021), demonstrates that contrastive learning with data augmentation can recover shared content without requiring such strong assumptions. This has been extended to MMCL by Lyu et al. (2022); Daunhawer et al. (2022), and Yao et al. Yao et al. (2023) further investigate identifiability under partial observability in multi-view settings. Distinct from prior work, we do not assume fixed content–style decompositions, causal directions, or perfectly aligned pairs. Instead, we analyze MMCL with image–text pairs under misalignment and systematically examine its impact on representation learning.

Vision-Language Models and Perspectives on Misalignment. Multimodal contrastive learning (MMCL) has achieved significant empirical success, particularly in aligning visual and textual modalities using models such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021). These successes are partly attributed to the use of massive training corpora, e.g., LAION-5B Schuhmann et al. (2022), which are substantially larger than those used for vision-only foundation models He et al. (2022); Oquab et al. (2024). However, real-world multimodal datasets are often imperfectly aligned and noisy Miech et al. (2019). Existing empirical methods Alayrac et al. (2020); Lavoie et al. (2024) typically treat such misalignment as label noise, employing strategies such as multiple-instance learning or dataset refinement Alayrac et al. (2020) to mitigate its impact. While some recent work suggests that contrastive models are robust to certain forms of structured misalignment Nakada et al. (2023), others propose augmenting text to simulate semantic variation in visual content Cai et al. (2025). Our work suggests that misalignment can act as either a barrier or a bridge, depending on the application. Unlike Nakada et al. (2023), which assumes linear representations without modeling the generative process, our analysis is grounded in a realistic latent variable model that provides a deeper understanding of misalignment from a data-generating perspective.

Identifiability in Latent Variable Models. Identifiability analysis addresses the fundamental question of whether the learning process can uniquely recover the latent generative structure or distribution underlying the observed data. This problem has been extensively studied in the context of nonlinear independent component analysis (ICA) (Hyvärinen and Pajunen, 1999; Locatello et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020) and causal representation learning (Liu et al., 2022, 2023; Zhang et al., 2024a; Yao et al., 2025; Liu et al., 2025). In practice, full identifiability—typically up to permutation—is rarely achievable without strong assumptions. Consequently, recent works have focused on partial identifiability (Gu and Xu, 2020; Kong et al., 2023; Gao et al., 2025) or relaxed equivalence classes, such as identifiability up to linear transformations (Zimmermann et al., 2021; Liu et al., 2024c) or up to group-wise/block-wise indeterminacy (Von Kügelgen et al., 2021; Yao et al., 2023), which can offer sufficient guarantees for specific tasks or settings. In the context of multimodal representation learning, several recent studies have explored identifiability results (Lyu et al., 2022; Daunhawer et al., 2022; Liu et al., 2024c), but largely neglect the presence of systematic misalignment. In contrast, our work explicitly models misalignment and adopts a block-identifiability definition to characterize the extent to which semantic factors can be recovered up to an invertible mapping.

Invariant Representation Learning. Invariant representation learning (IRL) seeks to learn representations that remain robust under distributional shifts between environments (Arjovsky et al., 2019; Eastwood et al., 2022; Gao et al., 2025), particularly in settings where empirical risk minimization (ERM) (Vapnik, 1991) fails to generalize out of distribution. In the absence of such variation, ERM is sufficient for in-distribution prediction, rendering the objective of IRL ill-posed. From a causal perspective, learning invariant representations—or more ambitiously, invariant mechanisms—requires variability in the non-invariant factors of the data. Such variability can be introduced through interventional data (Lippe et al., 2022; Lachapelle et al., 2022), exchangeability assumptions (Reizinger et al., 2025), or the use of auxiliary variables such as domain indices (Zhang et al., 2024a; Liu et al., 2024b). However, direct interventions on latent variables are typically infeasible in real-world data, and auxiliary variable methods often require access to a large number of diverse environments to ensure identifiability—an assumption that is rarely satisfied in practice. Our work offers an alternative approach: by leveraging the inherent flexibility of text supervision, we demonstrate that manipulating biases—specifically through selective omission or semantic perturbation in captions—can serve as a controllable proxy for environmental variation.

7 Conclusion and Discussion

In this work, we present a formal analysis of misalignment in MMCL, examining its impact on learned representations. We demonstrate that contrastive multimodal encoders retain only the unbiased shared semantic variables, systematically discarding misaligned latent variables. When image-text pairs exhibit selection or perturbation biases, the joint embedding prioritizes consistent content, while omitting altered or missing aspects. This trade-off is fundamental: perfectly aligned text captions preserve rich semantic detail, whereas selective or biased text can enhance domain invariance by filtering out distribution-sensitive factors. Our experiments, conducted across simulations and image-text datasets, empirically validate these theoretical findings. These insights underscore the need for multimodal learning frameworks that either mitigate misalignment or leverage beneficial biases to improve representation learning in real-world settings (see App. G for further discussion).

References

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. https://dl.acm.org/doi/10.5555/3495724.3495727.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. https://arxiv.org/abs/1907.02893.

V.I. Bogachev. *Measure Theory*. Number Volume I in Measure Theory. Springer Berlin Heidelberg, 2007. ISBN 9783540345145. https://books.google.com.au/books?id=CoSIe7h5mTsC.

Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi. Clap: Isolating content from style through contrastive learning with augmented prompts. In *European Conference on Computer Vision*, pages 130–147. Springer, 2025. https://doi.org/10.1007/978-3-031-72664-4_8.

Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE transactions on medical imaging*, 40(3):781–792, 2020. https://doi.org/10.1109/TMI.2020.3036584.

Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26836–26845, 2024. https://doi.org/10.1109/CVPR52733.2024.02534.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. http://www.blender.org.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 2006. https://doi.org/10.1002/047174882X.

George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951. https://doi.org/10.2307/1401511.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=U_2kuqoTcB.

Ning Ding, Chaorui Deng, Mingkui Tan, Qing Du, Zhiwei Ge, and Qi Wu. Image captioning with controllable and adaptive length levels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. https://ieeexplore.ieee.org/document/10310015.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022. https://dl.acm.org/doi/10.5555/3600270.3601531.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Fang_From_Captions_to_2015_CVPR_paper.pdf.

Erdun Gao, Howard Bondell, Shaoli Huang, and Mingming Gong. Domain generalization via content factors isolation: a two-level latent variable modeling approach. *Machine Learning*, 114(88), 2025. https://link.springer.com/article/10.1007/s10994-024-06717-6.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

Yuqi Gu and Gongjun Xu. Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4):2082–2107, 2020. https://doi.org/10.1214/19-AOS1878.

A. Hatcher. *Algebraic Topology*. Algebraic Topology. Cambridge University Press, 2002. ISBN 9780521795401. https://books.google.com.au/books?id=DFGU6HxQIVUC.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. https://doi.org/10.1109/CVPR52688.2022.01553.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. https://doi.org/10.1016/S0893-6080(98)00140-3.

Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. https://doi.org/10.1109/PROC.1982.12425.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. https://proceedings.mlr.press/v139/jia21b.html.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. https://proceedings.mlr.press/v108/khemakhem20a.

Bumsoo Kim, Yeonsik Jo, Jinhyung Kim, and Seunghwan Kim. Misalign, contrast then distill: Rethinking misalignments in language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2563–2572, October 2023. https://openaccess.thecvf.com/content/ICCV2023/papers/Kim_Misalign_Contrast_then_Distill_Rethinking_Misalignments_in_Language-Image_Pre-training_ICCV_2023_paper.pdf.

Bumsoo Kim, Jinhyung Kim, Yeonsik Jo, and Seung Hwan Kim. Expediting contrastive language-image pretraining via self-distilled encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2732–2740, 2024. https://doi.org/10.1609/aaai.v38i3.28052.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial identifiability for domain adaptation. *arXiv preprint arXiv:2306.06510*, 2023. https://proceedings.mlr.press/v162/kong22a.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. https://proceedings.mlr.press/v139/krueger21a/krueger21a.pdf.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022. https://openreview.net/forum?id=dHsFFekd_-o.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. *Proceedings of Machine Learning Research*, 235:26070–26084, 2024. https://proceedings.mlr.press/v235/lavoie24a.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. https://github.com/DAMO-NLP-SG/VCD.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 2022. https://proceedings.mlr.press/v162/li22n/li22n.pdf.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10421–10434. Curran Associates, Inc., 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/43d2b7fbee8431f7cef0d0afed51c691-Paper-Conference.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13557–13603. PMLR, 17–23 Jul 2022. `https://proceedings.mlr.press/v162/lippe22a.html`.

Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1503–1512, 2024a. `https://doi.org/10.1145/3627673.3679626`.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022. `https://arxiv.org/abs/2208.14153`.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. *arXiv preprint arXiv:2310.15580*, 2023. `https://openreview.net/forum?id=ia9fKO1Vjq&noteId=lUddseJaKf`.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent neural causal models. *arXiv preprint arXiv:2403.15711*, 2024b. `https://arxiv.org/pdf/2403.15711`.

Yuhang Liu, Zhen Zhang, Dong Gong, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Revealing multimodal contrastive representation learning through latent partial causal models. *arXiv preprint arXiv:2402.06223*, 2024c. `https://arxiv.org/abs/2402.06223`.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Latent covariate shift: Unlocking partial identifiability for multi-source domain adaptation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. `https://openreview.net/forum?id=9kFl0yLwyf`.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. `https://proceedings.mlr.press/v97/locatello19a.html`.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2554–2562, 2021. `https://doi.org/10.1109/CVPR46437.2021.00258`.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022. `https://openreview.net/forum?id=5FUq05QRc5b`.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. `https://www.di.ens.fr/willow/research/howto100m/`.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380. PMLR, 2023. `https://proceedings.mlr.press/v206/nakada23a.html`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. `https://arxiv.org/abs/1807.03748`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. `https://openreview.net/forum?id=a68SUt6zFt`.

Judea Pearl. *Causality*. Cambridge university press, 2009. `https://bayes.cs.ucla.edu/BOOK-2K/`.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. https://doi.org/10.1111/rssb.12167.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. https://proceedings.mlr.press/v139/radford21a.

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=k03mB41vyM.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. https://jmlr.org/papers/v19/16-432.html.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN 9780070856134. https://books.google.com.au/books?id=kwqzPAAACAAJ.

Michael Ruzhansky and Mitsuru Sugimoto. On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences*, 5:13–18, 2015. https://doi.org/10.1007/s13373-014-0059-1.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. https://aclanthology.org/P18-1238/".

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=rygeHgSFDH.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Annual Meeting of the Association for Computational Linguistics*, 2024. https://aclanthology.org/2024.findings-acl.775.pdf.

Tang Tao, Guangrun Wang, Yixing Lao, Peng Chen, Jie Liu, Liang Lin, Kaicheng Yu, and Xiaodan Liang. Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21230–21240, 2024.

V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. https://openreview.net/forum?id=4pf_pOo0Dt.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024. https://proceedings.neurips.cc/paper_files/paper/2023/file/97fe251c25b6f99a2a23b330a75b11d4-Paper-Conference.pdf.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. https://proceedings.mlr.press/v119/wang20k.

Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press, 2012.

Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, and Linjie Yang. Why is prompt tuning for vision-language models robust to noisy labels? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15497, 2023. https://doi.org/10.1109/ICCV51070.2023.01420.

Junru Wu, Yi Liang, Hassan Akbari, Zhangyang Wang, Cong Yu, et al. Scaling multimodal pretraining via cross-modality gradient harmonization. *Advances in Neural Information Processing Systems*, 35:36161–36173, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/eacad5b8e67850f2b8dd33d87691d097-Abstract-Conference.html.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63529–63541. Curran Associates, Inc., 2023. https://proceedings.neurips.cc/paper_files/paper/2023/file/c89f09849eb5af489abb122394ff0f0b-Paper-Conference.pdf.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273, 2024. https://aclanthology.org/2024.findings-emnlp.775.pdf.

Hu Xu, Po-Yao Huang, Xiaoqing Tan, Ching-Feng Yeh, Jacob Kahn, Christine Jou, Gargi Ghosh, Omer Levy, Luke Zettlemoyer, Wen-tau Yih, et al. Altogether: Image captioning via re-aligning alt-text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19302–19318, 2024. https://aclanthology.org/2024.emnlp-main.1075.pdf.

Kai Xuan, Lei Xiang, Xiaoqian Huang, Lichi Zhang, Shu Liao, Dinggang Shen, and Qian Wang. Multimodal mri reconstruction assisted with spatial alignment network. *IEEE Transactions on Medical Imaging*, 41(9):2499–2509, 2022. https://doi.org/10.1109/TMI.2022.3164050.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023. https://openreview.net/forum?id=OGtnhKQJms.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=lk2Qk5xjeu.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60057–60075. PMLR, 21–27 Jul 2024a. https://proceedings.mlr.press/v235/zhang24br.html.

Yanan Zhang, Jiangmeng Li, Lixiang Liu, and Wenwen Qiang. Rethinking misalignment in vision-language model adaptation from a causal perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. https://openreview.net/forum?id=vwgWbCxeAQ.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. https://proceedings.mlr.press/v182/zhang22a.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. https://proceedings.mlr.press/v139/zimmermann21a.

NEGATE OR EMBRACE: ON HOW MISALIGNMENT SHAPES MULTIMODAL REPRESENTATION LEARNING

(APPENDIX)

Table of Contents

A Notation and Terminology

Table 1 provides a summary of the notations and terminologies used throughout the paper.

Table 1: **Notations and Terminologies used Throughout the paper.**

Observation and Latent Spaces	
$\mathcal{X}$	Image observation space ($\subseteq \mathbb{R}^{d_x}$)
$\mathcal{T}^{(\theta)}$	Text observation subspace under selection bias θ
$\mathcal{S}$	Latent semantic space ($\subseteq \mathbb{R}^{n_s}$)
$\mathcal{M}_x$	Image-specific non-semantic latent space
$\mathcal{M}_t$	Text-specific non-semantic latent space
$\mathbb{I}_{\mathbf{s}}$	Index set of semantic variables: $\{1, \ldots, n_s\}$
$\mathbb{I}_{\text{inv}}$	Index subset of semantic variables that remain invariant under distribution shift
$\mathbb{I}_{\text{var}}$	Index subset of semantic variables that vary under distribution shift
Mappings and Functions	
$\mathbf{g}_x$	Generative mapping for images: $\mathcal{S} \times \mathcal{M}_x \to \mathcal{X}$
$\mathbf{g}_{t^{(\theta)}}$	Generative mapping for text under selection bias θ: $\mathcal{S}_{\mathbb{I}_\theta} \times \mathcal{M}_t \to \mathcal{T}^{(\theta)}$
$\mathbf{f}_x$	Image encoder: $\mathcal{X} \to (0,1)^n$ with specified n
$\mathbf{f}_t$	Text encoder: $\mathcal{T}^{(\theta)} \to (0,1)^n$ with specified n
Loss Functions	
$\mathcal{L}_{\text{MMCL}}(\mathbf{f}_x, \mathbf{f}_t)$	Symmetric InfoNCE loss for MMCL (Eq. (1))
$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x, \mathbf{f}_t)$	Alignment and entropy maximization loss (Eq. (2))
Notations for Misalignment	
θ	Selection bias, an integer realization in the range $\{1, \ldots, 2^{n_s} - 1\}$
$\mathcal{P}_+(\mathbb{I}_{\mathbf{s}})$	The set of all non-empty subsets of $\mathbb{I}_{\mathbf{s}}$
$\mathbb{I}_\theta$	A selected semantic subset indexed by θ, $\mathbb{I}_\theta \in \mathcal{P}_+(\mathbb{I}_{\mathbf{s}})$
$\mathbb{I}_\theta^c$	Omitted semantic subset under θ, $\mathbb{I}_\theta^c = \mathbb{I}_{\mathbf{s}} \setminus \mathbb{I}_\theta$
ρ	Perturbation bias, an integer realization in the range $\{1, \ldots, 2^{\lvert\mathbb{I}_\theta\rvert} - 1\}$
$\mathcal{P}_{\text{proper}}(\mathbb{I}_\theta)$	The set of all proper subsets of $\mathbb{I}_\theta$
$\mathbb{I}_\rho$	A subset of $\mathbb{I}_\theta$ subject to perturbation indexed by ρ, $\mathbb{I}_\rho \in \mathcal{P}_{\text{proper}}(\mathbb{I}_\theta)$
$\mathbb{I}_\rho^c$	Subset of $\mathbb{I}_\theta$ that always be unbiased under ρ, $\mathbb{I}_\rho^c = \mathbb{I}_\theta \setminus \mathbb{I}_\rho$
$p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta} \mid \mathbf{s}, \theta, \rho}$	Perturbation conditional distribution, reflecting misalignment (Eq. (5))
$A \subseteq \mathbb{I}_\rho$	A random subset of semantic variables subject to perturbation, drawn from p_A
Distributions and Operators	
$p_{\mathbf{s}}$, $p_{\mathbf{m}_x}$, $p_{\mathbf{m}_t}$	Distributions over semantic, image-specific and text-specific variables, respectively
$H(\cdot)$	Differential entropy
$\delta(\cdot)$	Dirac delta function
Random Variables	
$\mathbf{x}$	Image observation sampled from $\mathcal{X}$
$\mathbf{t}^{(\theta)}$	Text observation under selection view θ, sampled from $\mathcal{T}^{(\theta)}$
$\mathbf{s}$	Latent semantic variables in $\mathcal{S}$
$\mathbf{s}_{\mathbb{I}_\theta}$	Selected latent semantic variables for generating text
$\mathbf{s}_{\mathbb{I}_\theta^c}$	Omitted latent semantic variables for generating text
$\mathbf{s}_{\mathbb{I}_\rho}$	Perturbable latent semantic variables for generating text
$\mathbf{s}_{\mathbb{I}_\rho^c}$	Unbiased latent semantic variables within the selected part
$\mathbf{m}_x$	Image-specific non-semantic variable in $\mathcal{M}_x$
$\mathbf{m}_t$	Text-specific non-semantic variable in $\mathcal{M}_t$
$\mathbf{z}_x$	Combined latent variable for images: $(\mathbf{s}, \mathbf{m}_x)$
$\mathbf{z}_{t^{(\theta)}}$	Combined latent variable for text view: $(\mathbf{s}_{\mathbb{I}_\theta}, \mathbf{m}_t)$

B Proofs

B.1 Lemmas

Before proceeding with the proof, we first establish the following lemmas.

Lemma B.1 (Global Minimum of $\mathcal{L}_{\text{SymAlignMaxEnt}}$)**.** *Under the assumptions of Thm. 4.1, the global minimum of*

$$\mathcal{L}_{\textit{SymAlignMaxEnt}}(\mathbf{f}_x, \mathbf{f}_t) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{t}^{(\theta)})\sim p_{\mathbf{x},\mathbf{t}^{(\theta)}}} [\|\mathbf{f}_x(\mathbf{x}) - \mathbf{f}_t(\mathbf{t}^{(\theta)})\|_2] - \frac{1}{2}\Big(H\big(\mathbf{f}_x(\mathbf{x})\big) + H\big(\mathbf{f}_t(\mathbf{t}^{(\theta)})\big)\Big), \tag{6}$$

is 0. This minimum can be attained by the following pair of smooth functions:

$$\mathbf{f}_x^* = \mathbf{d} \circ (\mathbf{g}_x^{-1})_{\mathbb{I}_\rho^c} : \mathcal{X} \to (0,1)^n, \tag{7}$$

$$\mathbf{f}_t^* = \mathbf{d} \circ (\mathbf{g}_{t^{(\theta)}}^{-1})_{\mathbb{I}_\rho^c} : \mathcal{T}^{(\theta)} \to (0,1)^n, \tag{8}$$

where:

- $\mathbf{g}_x$ *and* $\mathbf{g}_{t^{(\theta)}}$ *denote the true underlying generative mappings for images and paired text, respectively, as described in § 3.*
- *The operator* $(\cdot)_{\mathbb{I}_\rho^c}$ *extracts the components corresponding to the preserved semantic variables (i.e., unaffected by the selection bias nor the perturbation bias), with* $n = \left|\mathbb{I}_\rho^c\right|$ *being their dimensionality.*
- $\mathbf{d} = (d_1, \ldots, d_n)$ *is defined via the Darmois construction (Darmois, 1951; Hyvärinen and Pajunen, 1999; Von Kügelgen et al., 2021), where for each* $i \in [n]$ *(we abbreviate* $\{1, \cdots, n\}$ *as* $[n]$ *for any integer* $n > 1$ *for simplicity),*

  $$d_i(\mathbf{s}_{\mathbb{I}_\rho^c}) = \text{CDF}_i\big(s_{\mathbb{I}_\rho^c,i} \mid \mathbf{s}_{\mathbb{I}_\rho^c,[i-1]}\big) = \mathbb{P}\Big(S_{\mathbb{I}_\rho^c,i} \le s_{\mathbb{I}_\rho^c,i} \,\Big|\, \mathbf{s}_{\mathbb{I}_\rho^c,[i-1]}\Big),$$

  with CDF_i *denoting the conditional cumulative distribution of* $s_{\mathbb{I}_\rho^c,i}$ *given* $\mathbf{s}_{\mathbb{I}_\rho^c,[i-1]}$.

Proof of Lem. B.1. We prove that the candidate functions $\mathbf{f}_x^*$ and $\mathbf{f}_t^*$ in Equations (7) and (8) yield $\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x^*, \mathbf{f}_t^*) = 0$. Substituting these candidate functions into the loss in Eq. (6), we have

$$\begin{aligned}\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x^*, \mathbf{f}_t^*) = &\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{t}^{(\theta)})\sim p_{\mathbf{x},\mathbf{t}^{(\theta)}}} \Big[\|\mathbf{f}_x^*(\mathbf{x}) - \mathbf{f}_t^*(\mathbf{t}^{(\theta)})\|_2\Big] \\ &- \frac{1}{2}\Big(H\big(\mathbf{f}_x^*(\mathbf{x})\big) + H\big(\mathbf{f}_t^*(\mathbf{t}^{(\theta)})\big)\Big).\end{aligned}$$

By the invertibility of the generative processes $\mathbf{g}_x$ and $\mathbf{g}_{t^{(\theta)}}$ (see § 3), we may change variables to express the expectation over the latent variables:

$$\begin{aligned}\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x^*, \mathbf{f}_t^*) = &\mathop{\mathbb{E}}_{(\mathbf{z}_x,\mathbf{z}_{t^{(\theta)}})\sim p_{\mathbf{z}_x,\mathbf{z}_{t^{(\theta)}}}} \Big[\|\mathbf{d}(\mathbf{s}_{\mathbb{I}_\rho^c}) - \mathbf{d}(\tilde{\mathbf{s}}_{\mathbb{I}_\rho^c})\|_2\Big] \\ &- \frac{1}{2}\Big(H\big(\mathbf{d}(\mathbf{s}_{\mathbb{I}_\rho^c})\big) + H\big(\mathbf{d}(\tilde{\mathbf{s}}_{\mathbb{I}_\rho^c})\big)\Big),\end{aligned}$$

where $\mathbf{s}_{\mathbb{I}_\rho^c}$ and $\tilde{\mathbf{s}}_{\mathbb{I}_\rho^c}$ denote the preserved unbiased components of the semantic variables across image-text pairs, respectively.

We now show that these unbiased semantic components are identical across modalities almost everywhere (a.e.). By Asm. 3.2, for any image-text pair the text is generated via a random perturbation process that modifies only a subset $A \subseteq \mathbb{I}_\rho$ of the activated semantic variables. Specifically, recall Eq. (5), the perturbation density is defined as

$$p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\big(\tilde{\mathbf{s}}_{\mathbb{I}_\theta} \mid \mathbf{s}, A\big) = \delta\big(\tilde{\mathbf{s}}_{\mathbb{I}_\theta\setminus A} - \mathbf{s}_{\mathbb{I}_\theta\setminus A}\big)\, p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}\big(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A\big).$$

Since $A \subseteq \mathbb{I}_\rho$, it follows that the indices in $\mathbb{I}_\rho^c$ are a subset of those in $\mathbb{I}_\theta \setminus A$; that is, $\mathbb{I}_\rho^c \subseteq \mathbb{I}_\theta \setminus A$. Thus, the Dirac delta in the above expression enforces that

$$\tilde{\mathbf{s}}_{\mathbb{I}_\rho^c} = \mathbf{s}_{\mathbb{I}_\rho^c} \quad \text{almost surely (a.s.)} \quad \forall\, \mathbf{s} \sim p_{\mathbf{s}}, \tilde{\mathbf{s}}_{\mathbb{I}_\theta} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s}_{\mathbb{I}_\theta},\rho},$$

under selection bias θ and perturbation ρ, regardless of the particular perturbation set A at each time.

Further, by the properties of the Darmois construction (Darmois, 1951), the mapping $\mathbf{d}$ transforms $\mathbf{s}_{\mathbb{I}_\rho^c}$ into a uniform distribution over $(0,1)^n$ (with $n = |\mathbb{I}_\rho^c|$). Since the uniform distribution is the unique maximum entropy (i.e., zero) distribution on a bounded domain (under no further moment constraints) (Jaynes, 1982; Cover, 2006), the entropy terms in the loss are maximized. In the formulation of $\mathcal{L}_{\text{SymAlignMaxEnt}}$, this maximal entropy precisely cancels any potential reduction in the loss, ensuring that

$$\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x^*, \mathbf{f}_t^*) = 0.$$

Therefore, the global minimum of $\mathcal{L}_{\text{SymAlignMaxEnt}}$ is achieved at 0 by the given function pairs $\mathbf{f}_x^*$ and $\mathbf{f}_t^*$, completing the proof. □

Lemma B.2 (Uniformizing Mapping Preserves All Information)**.** *Let $\mathbf{h} : \mathcal{U} \to \mathcal{V}$ be a smooth map between simply connected, open $\mathcal{C}^1$ manifolds $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^n$. Suppose that $\mathbf{u}$ is a random variable taking values in $\mathcal{U}$ with a smooth probability density that is strictly positive a.e.. If the pushforward $\mathbf{v} = \mathbf{h}(\mathbf{u})$ is uniformly distributed on $\mathcal{V}$, then $\mathbf{h}$ is a global diffeomorphism; in particular, $\mathbf{h}$ is bijective and depends on every component of $\mathbf{u}$.*[5]

Proof of Lem. B.2. Let $p_{\mathbf{u}} : \mathcal{U} \to \mathbb{R}$ and $p_{\mathbf{v}} : \mathcal{V} \to \mathbb{R}$ denote the probability density functions of $\mathbf{u}$ and $\mathbf{v}$, respectively. Since $\mathbf{v} = \mathbf{h}(\mathbf{u})$, the change-of-variables formula (refer to, e.g., Bogachev, 2007) yields

$$p_{\mathbf{v}}(\mathbf{v}) = p_{\mathbf{u}}(\mathbf{u}) \cdot \Big|\det J(\mathbf{h})(\mathbf{u})\Big|^{-1},$$

where $\mathbf{u}$ is any preimage of $\mathbf{v}$ under $\mathbf{h}$. By assumption, $\mathbf{v}$ is uniformly distributed on $\mathcal{V}$; that is, there exists a constant $C > 0$ such that

$$p_{\mathbf{v}}(\mathbf{v}) = C \quad \text{for all } \mathbf{v} \in \mathcal{V}.$$

Thus, for any $\mathbf{u}$ with $\mathbf{v} = \mathbf{h}(\mathbf{u})$ we obtain

$$\Big|\det J(\mathbf{h})(\mathbf{u})\Big|^{-1} = \frac{C}{p_{\mathbf{u}}(\mathbf{u})}.$$

Since $p_{\mathbf{u}}(\mathbf{u})$ is strictly positive a.e. on $\mathcal{U}$, it follows that

$$\Big|\det J(\mathbf{h})(\mathbf{u})\Big|^{-1} > 0 \quad \text{a.s.} \quad \forall\, \mathbf{u} \sim p_{\mathbf{u}},$$

or equivalently, $\det J(\mathbf{h})(\mathbf{u}) \neq 0$ a.e.. By the Inverse Function Theorem (see, e.g., Rudin, 1976), this implies that $\mathbf{h}$ is a local diffeomorphism.

Moreover, since $\mathcal{U}$ and $\mathcal{V}$ are simply connected, open $\mathcal{C}^1$ manifolds, standard covering space theory (refer to, e.g., the discussion around Theorem 1.38 in Hatcher, 2002) implies that $\mathbf{h}$ is a covering map. The uniformity of $p_{\mathbf{v}}$ forces $\mathbf{h}$ to be surjective (otherwise, some points in $\mathcal{V}$ would have zero density, contradicting uniformity). Since any covering map from a simply connected space is trivial, $\mathbf{h}$ must be equivalent to the identity covering. In other words, $\mathbf{h}$ is a homeomorphism onto $\mathcal{V}$ and hence both injective and surjective (i.e., a bijection).

Finally, the fact that the Jacobian determinant is nonzero a.e. guarantees that $\mathbf{h}$ depends on all components of $\mathbf{u}$; if any component were omitted, the rank of the Jacobian would drop, contradicting non-singularity. Furthermore, by the Global Inverse Function Theorem (refer to, e.g., Ruzhansky and Sugimoto, 2015), the inverse of $\mathbf{h}$ is smooth.

In summary, $\mathbf{h}$ is a global diffeomorphism from $\mathcal{U}$ onto $\mathcal{V}$. Consequently, it preserves all information of $\mathbf{u}$: every variation in $\mathbf{u}$ is reflected in $\mathbf{v} = \mathbf{h}(\mathbf{u})$, and $\mathbf{u}$ can be uniquely and smoothly recovered from $\mathbf{v}$. This completes the proof. □

B.2 Proof of Theorem 4.1

We now proceed to prove Thm. 4.1. To begin, we restate the theorem for clarity:

[5] We do not claim originality for this result due to its fundamental nature in topology and measure theory; rather, we detail it here as a tool for our subsequent arguments.

Theorem 4.1 (Identifiability of Latent Semantic Variables)**.** *Let $(\mathbf{x}, \mathbf{t}^{(\theta)})$ be image-text pairs drawn from the data-generating process described in § 3, where $\mathbf{x}$ is generated according to Eq.* (3) *and $\mathbf{t}_\theta$ is generated by Eq.* (4)*. Suppose that Asms. 3.1 and 3.2 hold. Denote by $\mathbf{s}_{\mathbb{I}_\rho^c}$ the subset of semantic variables that preserved without bias in the text, and define its dimension as $n = |\mathbb{I}_\theta| - |\mathbb{I}_\rho|$. Let $\mathbf{f}_x : \mathcal{X} \to (0,1)^n$ and $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^n$ be sufficiently flexible, smooth functions. Then, Minimizing the loss $\mathcal{L}_{\mathit{SymAlignMaxEnt}}$ in Eq.* (2) *over samples $(\mathbf{x}, \mathbf{t}_\theta)$ guarantees that $\mathbf{f}_x$ and $\mathbf{f}_t$ block-identify the semantic variables $\mathbf{s}_{\mathbb{I}_\rho^c}$ in the sense of Defn. 4.1.*

Proof of Thm. 4.1*.* The proof is organized into the following five steps:

1. First, we show that the objective function $\mathcal{L}_{\text{SymAlignMaxEnt}}(\mathbf{f}_x, \mathbf{f}_t)$ achieves a global minimum value of 0. At this minimum, any pair of smooth functions $\mathbf{f}_x$ and $\mathbf{f}_t$ satisfying this condition must exhibit invariance across modalities. This invariance condition ensures that the learned image representations and text representations must align across all positive $\mathbf{x}$ and $\mathbf{t}^{(\theta)}$ pairs.

2. Next, we prove that minimizing $\mathcal{L}_{\text{SymAlignMaxEnt}}$ inherently eliminates any dependence of the learned representations on modality-specific variables $\mathbf{m}_x$ or $\mathbf{m}_t$. This ensures that the representations are restricted to the dependence on latent semantic variables.

3. By contradiction, we further establish that any contribution from the omitted semantic variables induced by selection bias θ, i.e., $\mathbf{s}_{\mathbb{I}_\theta^c}$, would violate the invariance condition established in Step 1. This guarantees that the representations exclude the dependence on omitted semantic variables.

4. We then establish the exclusion of perturbed semantic variables influenced by perturbation bias ρ, i.e., $\mathbf{s}_{\mathbb{I}_\rho}$, from the learned representations, also by contradiction.

5. Finally, we demonstrate that the optimized mappings are invertible with respect to the learned representations and the true unbiased semantic variables $\mathbf{s}_{\mathbb{I}_\rho^c}$. This ensures that the representations block-identify the preserved unbiased semantic variables, thereby concluding the proof.

Step 1 (Global Minimum and Invariance Condition)**.** Let $\mathbf{f}_x : \mathcal{X} \to (0,1)^n$ and $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^n$ be any smooth functions attaining the global minimum. Define the smooth mappings:

$$\mathbf{h}_x = \mathbf{f}_x \circ \mathbf{g}_x, \quad \mathbf{h}_t = \mathbf{f}_t \circ \mathbf{g}_{t^{(\theta)}}.$$

Since all terms in $\mathcal{L}_{\text{SymAlignMaxEnt}}$ are non-negative, and its global minimum is 0 by Lem. B.1, each term in $\mathcal{L}_{\text{SymAlignMaxEnt}}$ must vanish a.s. for any pairing $(\mathbf{x}, \mathbf{t}^{(\theta)})$, leading to:

$$\underset{(\mathbf{x},\mathbf{t}^{(\theta)})\sim p_{\mathbf{x}}\, p_{\mathbf{t}^{(\theta)}|\mathbf{x}}}{\mathbb{E}} \left[\|\mathbf{f}_x(\mathbf{x}) - \mathbf{f}_t(\mathbf{t}^{(\theta)})\|_2\right] = \underset{(\mathbf{z}_x,\mathbf{z}_{t^{(\theta)}})\sim p_{\mathbf{z}_x}\, p_{\mathbf{z}_{t^{(\theta)}}|\mathbf{z}_x}}{\mathbb{E}} \left[\|\mathbf{h}_x(\mathbf{z}_x) - \mathbf{h}_t(\mathbf{z}_{t^{(\theta)}})\|_2\right] = 0, \tag{9}$$

$$H\big(\mathbf{f}_x(\mathbf{x})\big) = H\big(\mathbf{h}_x(\mathbf{z}_x)\big) = 0, \tag{10}$$

$$H\big(\mathbf{f}_t(\mathbf{t}^{(\theta)})\big) = H\big(\mathbf{h}_t(\mathbf{z}_{t^{(\theta)}})\big) = 0. \tag{11}$$

From Eq. (9), it follows that

$$\mathbf{h}_t(\mathbf{z}_{t^{(\theta)}}) = \mathbf{h}_x(\mathbf{z}_x), \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_{\mathbf{s}}\, p_{\mathbf{m}_x},\ \mathbf{z}_{t^{(\theta)}} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\, p_{\mathbf{m}_t}, \tag{12}$$

which ensures alignment of representations a.s. for any pair $(\mathbf{x}, \mathbf{t}^{(\theta)})$. The substitution of expectations for the image modality in Eq. (9) is valid because $\mathbf{x}$ follows the pushforward distribution of $\mathbf{z}_x$ under the deterministic diffeomorphism $\mathbf{g}_x$ (Eq. (3)); a similar argument applies to the text (Eq. (4)).

Equations (10) and (11) imply that $\mathbf{h}_x$ and $\mathbf{h}_t$ map the latent variables $\mathbf{z}_x$ and $\mathbf{z}_{t^{(\theta)}}$ onto uniform distributions over $(0,1)^n$ (with $n = |\mathbb{I}_\rho^c|$), since their differential entropy equals to zero.

Step 2 (Exclusion of Modality-Specific Variables) **.** We now show that the smooth functions $\mathbf{h}_x$ and $\mathbf{h}_t$ depend only on latent semantic variables $\mathbf{s}$ (with further exclusion of components in later steps) and not on modality-specific variables $\mathbf{m}_x$ or $\mathbf{m}_t$.

Since $\mathbf{z}_x = (\mathbf{s}, \mathbf{m}_x)$ and $\mathbf{z}_{t^{(\theta)}} = (\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t)$ are the latent variables generating images $\mathbf{x}$ and paired text $\mathbf{t}^{(\theta)}$, respectively, the assumed data-generating process (§ 3) implies the following independence properties:

(**c1**) $\mathbf{m}_x$ is independent of $\mathbf{z}_{t(\theta)}$: This means changes in $\mathbf{m}_x$ do not influence $\mathbf{z}_{t(\theta)}$. Moreover, since the text generation process $\mathbf{g}_{t(\theta)}$ is independent of $\mathbf{m}_x$, it follows that:

$$\mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t, \mathbf{m}_x) = \mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t). \tag{13}$$

(**c2**) $\mathbf{m}_t$ is independent of $\mathbf{z}_x$: This means changes in $\mathbf{m}_t$ do not influence $\mathbf{z}_x$. Similarly, since the image generation process $\mathbf{g}_x$ is independent of $\mathbf{m}_t$, it follows that:

$$\mathbf{h}_x(\mathbf{s}, \mathbf{m}_x, \mathbf{m}_t) = \mathbf{h}_x(\mathbf{s}, \mathbf{m}_x). \tag{14}$$

Combining Equations (12) and (13), we have:

$$\mathbf{h}_x(\mathbf{s}, \mathbf{m}_x) = \mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t, \mathbf{m}_x), \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_{\mathbf{s}}\, p_{\mathbf{m}_x},\ \mathbf{z}_{t(\theta)} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\, p_{\mathbf{m}_t}. \tag{15}$$

Consider a small perturbation $\mathbf{m}_x \mapsto \mathbf{m}_x + \boldsymbol{\varsigma}$, where $\|\boldsymbol{\varsigma}\|$ is arbitrarily small but remains within the open space $\mathcal{M}_x$. Since changes in $\mathbf{m}_x$ do not influence $\mathbf{h}_t$ by statement (**c1**), we obtain:

$$\mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t, \mathbf{m}_x + \boldsymbol{\varsigma}) = \mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t, \mathbf{m}_x). \tag{16}$$

Since $\mathbf{m}_x$ is independent of $\mathbf{s}$, perturbations in $\mathbf{m}_x$ does not alter the semantic variables, and $p_{\mathbf{m}_x} > 0$ a.e. over $\mathcal{M}_x$ by Asm. 3.1. Thus, substituting

$$(\mathbf{s}, \mathbf{m}_x) \mapsto (\mathbf{s}, \mathbf{m}_x + \boldsymbol{\varsigma})$$

in Eq. (15) and combining with Eq. (16), we get:

$$\mathbf{h}_x(\mathbf{s}, \mathbf{m}_x + \boldsymbol{\varsigma}) = \mathbf{h}_x(\mathbf{s}, \mathbf{m}_x).$$

By the smoothness of $\mathbf{h}_x$ (inherited from the smoothness of $\mathbf{g}_x$ and $\mathbf{f}_x$), taking $\boldsymbol{\varsigma} \to \mathbf{0}$ gives:

$$\frac{\partial \mathbf{h}_x}{\partial \mathbf{m}_x} = \lim_{\boldsymbol{\varsigma}\to\mathbf{0}} \frac{\mathbf{h}_x(\mathbf{s}, \mathbf{m}_x + \boldsymbol{\varsigma}) - \mathbf{h}_x(\mathbf{s}, \mathbf{m}_x)}{\boldsymbol{\varsigma}} = 0.$$

Thus, $\mathbf{h}_x$ is independent of $\mathbf{m}_x$.

A symmetric argument applies to $\mathbf{m}_t$. If $\mathbf{m}_t \mapsto \mathbf{m}_t + \boldsymbol{\varsigma}$, then by Eq. (14) in statement (**c2**), $\mathbf{h}_x(\mathbf{s}, \mathbf{m}_x, \mathbf{m}_t + \boldsymbol{\varsigma})$ remains unchanged. The invariance condition in Eq. (12) then forces $\mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}, \mathbf{m}_t + \boldsymbol{\varsigma})$ to remain constant w.r.t. $\boldsymbol{\varsigma}$, showing that $\mathbf{h}_t$ is independent of $\mathbf{m}_t$. Therefore, the learned representations satisfy:

$$\mathbf{h}_x(\mathbf{z}_x) = \mathbf{h}_x(\mathbf{s}), \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_{\mathbf{s}}\, p_{\mathbf{m}_x}, \tag{17}$$

$$\mathbf{h}_t(\mathbf{z}_{t(\theta)}) = \mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}), \quad \text{a.s.} \quad \forall\, \mathbf{z}_{t(\theta)} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\, p_{\mathbf{m}_t}. \tag{18}$$

Step 3 (Exclusion of Omitted Semantic Variables)**.** We now establish that the function $\mathbf{h}_x$ is independent of $\mathbf{s}_{\mathbb{I}_\theta^c}$, where $\mathbb{I}_\theta^c = \mathbb{I}_{\mathbf{s}} \setminus \mathbb{I}_\theta$. In other words, the learned representations do not contain information about the omitted semantic variables that are absent in the corresponding text.

Using the invariant condition in Eq. (12), together with the independence of modality-specific non-semantic variables in Equations (17) and (18), we have the following updated invariant condition:

$$\mathbf{h}_x(\mathbf{s}) \;=\; \mathbf{h}_t\big(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}\big), \quad \text{a.s.} \quad \forall\, \mathbf{s} \sim p_{\mathbf{s}},\ \tilde{\mathbf{s}}_{\mathbb{I}_\theta} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}. \tag{19}$$

Next, we show by contradiction that $\mathbf{h}_x$ is independent of $\mathbf{s}_{\mathbb{I}_\theta^c}$. Suppose, for the sake of contradiction, that there exists a function $\mathbf{h}_x^c = \mathbf{f}_x^c \circ \mathbf{g}_x$ which depends on at least one component of the omitted semantic variables $\mathbf{s}_{\mathbb{I}_\theta^c}$. Formally,

$$\exists\, l \in \mathbb{I}_\theta^c, \quad (\mathbf{s}^*_{\mathbb{I}_\theta}, \mathbf{s}^*_{\mathbb{I}_\theta^c}) \in \mathcal{S}, \quad \text{such that} \quad \frac{\partial\, \mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta}, \mathbf{s}^*_{\mathbb{I}_\theta^c}\big)}{\partial\, s_l^*} \;\neq\; 0.$$

By the C^1 continuity of $\mathbf{h}_x^c$, guaranteed by the smoothness of both $\mathbf{g}_x$ and $\mathbf{f}_x^c$, and that $\mathcal{S}$ is an open space, it follows that

$$\exists\, \eta > 0: \quad s_l \;\mapsto\; \mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta}, (s_l, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big) \quad \text{is strictly monotonic on} \quad (s_l^* - \eta,\ s_l^* + \eta),$$

where $\mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}}$ denotes all components of $\mathbf{s}^*_{\mathbb{I}_\theta^c}$ except s_l.

Since $p_{\mathbf{s}} > 0$ a.e. on $\mathcal{S}$, we can find two distinct latent semantic variables

$$\big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^-, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big),\ \big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^+, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big) \quad \text{with } s_l^- \in (s_l^* - \eta,\ s_l^*),\ s_l^+ \in (s_l^*,\ s_l^* + \eta) \tag{20}$$

that correspond to two different image observations, such that

$$\mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^-, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big) \;\neq\; \mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^+, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big). \tag{21}$$

However, combining Eq. (19), we have

$$\mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^-, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big) \;=\; \mathbf{h}_x^c\big(\mathbf{s}^*_{\mathbb{I}_\theta},\, (s_l^+, \mathbf{s}^*_{\mathbb{I}_\theta^c\setminus\{l\}})\big) \;=\; \mathbf{h}_t\big(\tilde{\mathbf{s}}^*_{\mathbb{I}_\theta}\big), \tag{22}$$

where $\tilde{\mathbf{s}}^*_{\mathbb{I}_\theta}$ represents the perturbed semantic variables of $\mathbf{s}^*_{\mathbb{I}_\theta}$ introduced by selection bias (with the exclusion of perturbed components further addressed below).

Equations (21) and (22) thus contradict each other. Hence, such a function $\mathbf{h}_x^c$ cannot exist. Consequently, $\mathbf{h}_x$ must be independent of $\mathbf{s}_{\mathbb{I}_\theta^c}$. Formally,

$$\mathbf{h}_x(\mathbf{z}_x) \;=\; \mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\theta}), \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_{\mathbf{s}}\, p_{\mathbf{m}_x}. \tag{23}$$

Clarification C.1 (Causal Interpretations)**. (i) Justification for the existence of distinct points in Eq. (20).** This follows from the assumption $p_{\mathbf{s}} > 0$ a.e. on $\mathcal{S}$ by Asm. 3.1. From a latent SCM perspective Pearl (2009); von Kügelgen et al. (2024), even if a specific semantic component s_l in $\mathbf{s}_{\mathbb{I}_\theta^c}$ is the ancestor node of some other semantic components in $\mathbf{s}_{\mathbb{I}_\theta}$, the strict positivity of $p_{\mathbf{s}}$ ensures that the exogenous noise variables are well-defined. Thus, for different values of s_l, there exist corresponding noise values that keep $\mathbf{s}^*_{\mathbb{I}_\theta}$ remaining fixed. **(ii) What if the unknown causal structure is $\mathbf{s}_{\mathbb{I}_\theta^c} \rightarrow \mathbf{s}_{\mathbb{I}_\theta}$?** The potential causal influence from $\mathbf{s}_{\mathbb{I}_\theta^c}$ to $\mathbf{s}_{\mathbb{I}_\theta}$ does not resolve the contradiction. Independence here means that, once $\mathbf{s}_{\mathbb{I}_\theta}$ is set, there is no direct functional path from $\mathbf{s}_{\mathbb{I}_\theta^c}$ to the representations $\mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\theta})$, i.e., the causal influence among them is fully accounted for by the realized value of $\mathbf{s}_{\mathbb{I}_\theta}$.

In summary, these arguments show that $\mathbf{h}_x$ is genuinely independent of $\mathbf{s}_{\mathbb{I}_\theta^c}$, even allowing for arbitrary unknown causal interactions among the latent semantic variables.

Step 4 (Exclusion of Perturbed Semantic Variables)**.** We now demonstrate that both representations are independent of $\mathbf{s}_{\mathbb{I}_\rho}$ and $\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$ respectively, as a consequence of the contradiction between the invariant condition and the random perturbations introduced by perturbation bias.

First, we refine the invariant condition by excluding omitted semantic variables as established above. Combining Equations (12), (18) and (23), we obtain:

$$\mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\theta}) \;=\; \mathbf{h}_t\big(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}\big), \quad \text{a.s.} \quad \forall\, \mathbf{s} \sim p_{\mathbf{s}},\ \tilde{\mathbf{s}}_{\mathbb{I}_\theta} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}. \tag{24}$$

Next, we show that $\mathbf{h}_t$ must be independent of $\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$ by contradiction. Suppose, for a contradiction, that there exist some function $\mathbf{h}_t^c = \mathbf{f}_t^c \circ \mathbf{g}_{t(\theta)}$ which depends on at least on component of the perturbed semantic variables $\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$. Formally,

$$\exists\, l \in \mathbb{I}_\rho, \quad (\tilde{\mathbf{s}}^*_{\mathbb{I}_\rho}, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c}), \quad \text{such that} \quad \frac{\partial \mathbf{h}_t^c(\tilde{\mathbf{s}}^*_{\mathbb{I}_\rho}, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c})}{\partial \tilde{s}_l^*} \neq 0.$$

By the C^1 continuity of $\mathbf{h}_t^c$ guaranteed by the smoothness of $\mathbf{f}_t^c$ and $\mathbf{g}_{t(\theta)}$, for some sufficiently small $\eta > 0$, we have the following inequality:

$$\mathbf{h}_t^c\big((s_l^-, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho\setminus\{l\}}), \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c}\big) \neq \mathbf{h}_t^c\big((s_l^+, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho\setminus\{l\}}), \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c}\big),\ \forall\, s_l^- \in (\tilde{s}_l^* - \eta, \tilde{s}_l^*),\ s_l^+ \in (\tilde{s}_l^*, \tilde{s}_l^* + \eta). \tag{25}$$

On the other hand, by the pairing conditions in Asm. 3.2, there exists at least one subset $A \subseteq \mathbb{I}_\rho$ of perturbed semantic variables such that $l \in A$ and $p_A(A) > 0$. Pick one such set and call it A. Define the latent semantic variables corresponding to the image of this pair as $(\mathbf{s}^*_A, \mathbf{s}^*_{\mathbb{I}_\theta\setminus A})$ (here, we omit the components that have already been excluded in previous steps).

By Asm. 3.2, we know that $\mathbf{s}^*_{\mathbb{I}_\theta\setminus A} = \tilde{\mathbf{s}}^*_{\mathbb{I}_\theta\setminus A}$ a.e., that is, $(\mathbf{s}_{\mathbb{I}_\rho\setminus A}, \mathbf{s}^*_{\mathbb{I}_\rho^c}) = (\tilde{\mathbf{s}}_{\mathbb{I}_\rho\setminus A}, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c})$ a.e.. Thus, we can rewrite text semantic variables $\tilde{\mathbf{s}}^*_{\mathbb{I}_\theta} = (\tilde{\mathbf{s}}^*_{\mathbb{I}_\rho}, \tilde{\mathbf{s}}^*_{\mathbb{I}_\rho^c})$ as $(\tilde{\mathbf{s}}^*_A, \mathbf{s}^*_{\mathbb{I}_\theta\setminus A})$.

Further, also by Asm. 3.2, there exists a non-empty open subspace $\mathcal{O}_A \subseteq \mathcal{S}_A$ such that $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A^*)$ is strictly positive on $\mathcal{O}_A$. Since the perturbed random variable $\tilde{\mathbf{s}}_A^*$ is a realization within this open subspace, we know it lies in $\mathcal{O}_A$ and $\mathcal{O}_A$ is non-empty. Moreover, because $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A^*)$ is smooth and strictly positive on $\mathcal{O}_A$, there exists a sufficiently small $\eta_1 > 0$ such that

$$p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A^*) > 0,\ \forall\, \tilde{\mathbf{s}}_A \in \{\tilde{\mathbf{s}}^*_{A\setminus\{l\}}\} \times (\tilde{s}_l^* - \eta_1, \tilde{s}_l^* + \eta_1),\ \text{with } \{\tilde{\mathbf{s}}^*_{A\setminus\{l\}}\} \times (\tilde{s}_l^* - \eta_1, \tilde{s}_l^* + \eta_1) \subseteq \mathcal{O}_A.$$

Thus, with a positive probability guaranteed by the above conditional, for the image semantic variables $(\mathbf{s}_A^*, \mathbf{s}^*_{\mathbb{I}_\theta\setminus A})$, we can construct two distinct realizations of perturbed semantic variables for generating different text:

$$\big((s_l^{(1)}, \tilde{\mathbf{s}}^*_{A\setminus\{l\}}), \mathbf{s}^*_{\mathbb{I}_\theta\setminus A}\big),\quad \big((s_l^{(2)}, \tilde{\mathbf{s}}^*_{A\setminus\{l\}}), \mathbf{s}^*_{\mathbb{I}_\theta\setminus A}\big),$$

where

$$s_l^{(1)} \in (\tilde{s}_l^* - \eta_2, \tilde{s}_l^*), \quad s_l^{(2)} \in (\tilde{s}_l^*, \tilde{s}_l^* + \eta_2) \quad \text{with} \quad \eta_2 = \min(\eta, \eta_1).$$

Based on the invariant condition established in Eq. (24), we have the following equalities:

$$\mathbf{h}_t^c\big((s_l^{(1)}, \tilde{\mathbf{s}}^*_{A\setminus\{l\}}), \mathbf{s}^*_{\mathbb{I}_\theta\setminus A}\big) = \mathbf{h}_t^c\big((s_l^{(2)}, \tilde{\mathbf{s}}^*_{A\setminus\{l\}}), \mathbf{s}^*_{\mathbb{I}_\theta\setminus A}\big) = \mathbf{h}_x(\mathbf{s}_A^*, \mathbf{s}^*_{\mathbb{I}_\theta\setminus A}).$$

This is contradicted by the inequality established in Eq. (25), which implies that such a $\mathbf{h}_t^c$ cannot exist. Consequently, any $\mathbf{h}_t$ minimizing the loss must be independent of the perturbed semantic variables $\tilde{\mathbf{s}}_{\mathbb{I}_\rho}$, i.e.,

$$\mathbf{h}_t(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}) = \mathbf{h}_t(\mathbf{s}_{\mathbb{I}_\rho^c}), \quad \text{a.s.} \quad \forall\, \tilde{\mathbf{s}}_{\mathbb{I}_\theta} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}. \tag{26}$$

Updating the invariant condition, and combining Equations (24) and (26), we obtain:

$$\mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\theta}) = \mathbf{h}_t(\mathbf{s}_{\mathbb{I}_\rho^c}), \quad \text{a.s.} \quad \forall\, \mathbf{s} \sim p_{\mathbf{s}}. \tag{27}$$

The exclusion of $\mathbf{s}_{\mathbb{I}_\rho}$ from image representations $\mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\theta})$ can be demonstrated by a similar procedure to Step 3, namely, that excluded semantic components from one modality cannot exist in another view, regardless of the latent causal structure among $\mathbf{s}_{\mathbb{I}_\theta}$. Specifically, fixing the value of $\mathbf{s}_{\mathbb{I}_\rho^c}$ and varying $\mathbf{s}_{\mathbb{I}_\rho}$ within a small region, we can sample distinct semantic variables (due to $p_{\mathbf{s}} > 0$ a.e. over $\mathcal{S}$ by Asm. 3.1). The smoothness of $\mathbf{h}_x$ then leads to an inequality if it depends on any component in $\mathbb{I}_\rho$. This inequality contradicts the alignment condition established in Eq. (27). Thus, $\mathbf{h}_x$ must also be independent of $\mathbf{s}_{\mathbb{I}_\rho}$.

Overall, due to the exclusion of modality-specific variables ($\mathbf{m}_x$ and $\mathbf{m}_t$), omitted semantic variables ($\mathbf{s}_{\mathbb{I}_\theta^c}$) and perturbed semantic variables ($\mathbf{s}_{\mathbb{I}_\rho}$) introduced by selection and perturbation biases for generating text, we now have the following equalities:

$$\mathbf{h}_x(\mathbf{z}_x) = \mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\rho^c}), \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_{\mathbf{s}}\, p_{\mathbf{m}_x}, \tag{28}$$

$$\mathbf{h}_t(\mathbf{z}_{t^{(\theta)}}) = \mathbf{h}_t(\mathbf{s}_{\mathbb{I}_\rho^c}), \quad \text{a.s.} \quad \forall\, \mathbf{z}_{t^{(\theta)}} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\, p_{\mathbf{m}_t}. \tag{29}$$

Step 5 (Preservation of All Unbiased Semantic Variables)**.** Based on Equations (28) and (29), we define the learned image and textual representations as

$$\hat{\mathbf{z}}_x = \mathbf{h}_x(\mathbf{s}_{\mathbb{I}_\rho^c}),\ \hat{\mathbf{z}}_t = \mathbf{h}_t(\mathbf{s}_{\mathbb{I}_\rho^c}), \quad \text{with} \quad \hat{\mathbf{z}}_x \in (0,1)^n,\ \hat{\mathbf{z}}_t \in (0,1)^n.$$

By directly applying Lem. B.2, it follows that the learned representations $\hat{\mathbf{z}}_x$ (and also $\hat{\mathbf{z}}_t$) include *all and only* the information of the unaltered semantic components $\mathbf{s}_{\mathbb{I}_\rho^c}$ almost surely, and that $\mathbf{h}_x$ (and similarly $\mathbf{h}_t$) is invertible. Consequently, the true modality-shared semantic variables $\mathbf{s}_{\mathbb{I}_\rho^c}$ are block-identified by $\mathbf{f}_x$ and $\mathbf{f}_t$.

Thus, the proof of Thm. 4.1 is complete. □

B.3 Proof of Corollary 4.1

We now proceed to prove Cor. 4.1. To begin, we restate the corollary for clarity:

Corollary 4.1 (Identifiability of Full Latent Semantic Variables)**.** *Let the selection bias be $\theta = 2^{n_s} - 1$ and the perturbation bias be $\rho = 1$, such that the full set of semantic variables $\mathbb{I}_{\mathbf{s}}$ is selected, and the perturbable semantic subset is trivial, i.e., $\mathbb{I}_\rho = \emptyset$. Then, all semantic variables $\mathbf{s}$ are block-identified via smooth functions $\mathbf{f}_x : \mathcal{X} \to (0,1)^{n_s}$ and $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^{n_s}$, when minimizing $\mathcal{L}_{\mathit{SymAlignMaxEnt}}$.*

Proof. As we have fixed a graded lexicographic order over the range of θ, that is, over $\mathcal{P}_+(\mathbb{I}_\mathbf{s})$ as defined in Defn. 3.1. Then, setting $\theta = 2^{n_s} - 1$ corresponds to selecting the full set of semantic variables for text generation, i.e., $\mathbb{I}_\theta = \mathbb{I}_\mathbf{s}$.

Furthermore, we have similarly fixed a graded lexicographic order over the range of ρ, i.e., over $\mathcal{P}_{\text{proper}}(\mathbb{I}_\mathbf{s})$, as defined in Defn. 3.2. Given that $\mathbb{I}_\theta = \mathbb{I}_\mathbf{s}$, setting $\rho = 1$ implies that all semantic variables $\mathbf{s}$ are preserved without perturbation during the generation of the corresponding text $\mathbf{t}^{(\theta)}$.

Under these assumptions, and by Asm. 3.2, the perturbing subset A is always trivial because $\mathbb{I}_\rho$ is trivial. Consequently, we have

$$p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}(\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s}) = \delta(\mathbf{s}-\mathbf{s}) \quad \text{with} \quad \mathbb{I}_\theta = \mathbb{I}_\mathbf{s},\ \mathbb{I}_\rho = \emptyset \quad \text{a.s.} \quad \forall\, \mathbf{s} \sim p_\mathbf{s},$$

which indicates that $\tilde{\mathbf{s}} = \mathbf{s}$ almost surely.

Now consider any pair of smooth functions $\mathbf{f}_x : \mathcal{X} \to (0,1)^{n_s}$ and $\mathbf{f}_{t^{(\theta)}} : \mathcal{T}^{(\theta)} \to (0,1)^{n_s}$ that achieve the global minimum of the loss in Eq. (2), i.e., yield a value of zero. From Step 1 and Step 2 of the proof of Thm. 4.1, it follows that:

$$\mathbf{h}_x(\mathbf{z}_x) = \mathbf{h}_t(\mathbf{z}_{t^{(\theta)}}) = \mathbf{h}_x(\mathbf{s}) = \mathbf{h}_t(\mathbf{s}) \quad \text{a.s.} \quad \forall\, \mathbf{z}_x \sim p_\mathbf{s}\, p_{\mathbf{m}_x},\ \mathbf{z}_{t^{(\theta)}} \sim p_{\tilde{\mathbf{s}}_{\mathbb{I}_\theta}|\mathbf{s},\theta,\rho}\, p_{\mathbf{m}_t}.$$

Since both the omitted and perturbable index subsets are trivial, we may directly apply Lem. B.2, which implies that the full semantic vector $\mathbf{s}$ is block-identified by both $\mathbf{f}_x$ and $\mathbf{f}_t$.

This completes the proof. □

Clarification C.2 (Fixed Graded Lexicographic Order)**.** The graded lexicographic order over index subsets of semantic variables is fixed purely for notational clarity and convenience. This choice imposes no constraints on the latent structure and is adopted without loss of generality. Since the true permutation of latent variables is unknown, any consistent ordering can be arbitrarily chosen to index the subsets associated with θ and ρ. Importantly, selection (θ) and perturbation (ρ) biases are applied indirectly at the level of text observations—not via direct interventions on the latent space. As a result, the ordering merely determines how each value of θ or ρ maps to a subset of semantic indices in $\mathbb{I}_\mathbf{s}$. For instance, omitting color information in captioning excludes the corresponding semantic variable; which value of θ encodes this depends on the chosen order. Fixing a graded lexicographic order therefore provides a deterministic and reproducible indexing scheme, without affecting the model's generality or expressiveness.

B.4 Proof of Corollary 4.2

We now proceed to prove Cor. 4.2. For clarity, we restate the corollary below:

Corollary 4.2 (Identifiability of Invariant Semantic Variables)**.** *Consider an OOD setting in which a subset of semantic variables, $\mathbb{I}_{inv} \subset \mathbb{I}_\mathbf{s}$, remains invariant between training and testing environments, while the remaining semantic variables, $\mathbb{I}_{var} = \mathbb{I}_\mathbf{s} \setminus \mathbb{I}_{inv}$, undergo distribution shifts. If the union of omitted and perturbable semantic variables under selection bias θ and perturbation bias ρ coincides with the environment-sensitive subset, i.e., $\mathbb{I}_{var} = \mathbb{I}_\theta^c \cup \mathbb{I}_\rho$, then the invariant semantic variables $\mathbf{s}_{\mathbb{I}_{inv}}$ are block-identified via smooth functions $\mathbf{f}_x : \mathcal{X} \to (0,1)^{|\mathbb{I}_{inv}|}$ and $\mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^{|\mathbb{I}_{inv}|}$, by minimizing $\mathcal{L}_{SymAlignMaxEnt}$.*

Proof. Under the OOD setting, and without loss of generality, let the subset of semantic variables susceptible to distribution shift be denoted by $\mathbb{I}_{\text{var}} = \mathbb{I}^1_{\text{var}} \cup \mathbb{I}^2_{\text{var}}$. Suppose the index set associated with selection bias is given by $\mathbb{I}_\theta = \mathbb{I}^1_{\text{var}} \cup \mathbb{I}_{\text{inv}}$, and that the perturbation bias acts on $\mathbb{I}_\rho = \mathbb{I}^1_{\text{var}}$. That is, the subset $\mathbb{I}^2_{\text{var}}$ is entirely omitted by the selection mechanism (i.e., excluded from $\mathbb{I}_\theta$), while the subset $\mathbb{I}^1_{\text{var}}$ is included but remains vulnerable to perturbation, as determined by ρ.

Given this structure, we directly apply the argument from the proof of Thm. 4.1. The omission of variables in $\mathbb{I}^2_{\text{var}}$, together with the perturbation of variables in $\mathbb{I}^1_{\text{var}}$, ensures that only the invariant subset $\mathbb{I}_{\text{inv}}$ is both selected and unperturbed. Therefore, the invariant semantic components $\mathbf{s}_{\mathbb{I}_{\text{inv}}}$ are block-identified via smooth functions

$$\mathbf{f}_x : \mathcal{X} \to (0,1)^{|\mathbb{I}_{\text{inv}}|} \quad \text{and} \quad \mathbf{f}_t : \mathcal{T}^{(\theta)} \to (0,1)^{|\mathbb{I}_{\text{inv}}|},$$

which attain the global minimum of the alignment objective.

This concludes the proof. □

C Numerical Simulation Details

We provide additional details on the numerical simulations that are not fully covered in § 5.1. Specifically:

- In App. C.1, we outline the experimental setup, including hyperparameters, model architectures, and the construction of downstream tasks.
- In App. C.2, we present additional experiments that further validate our theoretical findings.
- In App. C.3, we analyze downstream performance under various perturbation bias settings.

C.1 Detailed Experimental Setup

Latent Space Construction. We decompose the latent space into a semantic subspace $\mathcal{S} \subseteq \mathbb{R}^{10}$ and two modality-specific latent subspaces, $\mathcal{M}_x \subseteq \mathbb{R}^5$ and $\mathcal{M}_t \subseteq \mathbb{R}^5$. Latent variables are sampled from multivariate Gaussian distributions: $\mathbf{s} \sim \mathcal{N}(0, \Sigma_{\mathbf{s}})$, $\mathbf{m}_x \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_x})$, and $\mathbf{m}_t \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_t})$.

In the *independent semantic variables* setting, we set $\Sigma_{\mathbf{s}} = I_{10}$, so that $\mathbf{s}$ follows a factorized standard Gaussian distribution, consistent with nonlinear ICA settings Khemakhem et al. (2020); Sorrenson et al. (2020). In the *dependent semantic variables* setting, we sample $\Sigma_{\mathbf{s}}$ from a Wishart distribution with an identity scale matrix and degrees of freedom equal to 10, introducing an unknown causal structure among semantic latent variables, as in prior causal representation learning works Von Kügelgen et al. (2021); Daunhawer et al. (2022); Yao et al. (2023).

For modality-specific variables, distinct covariance matrices $\Sigma_{\mathbf{m}_x}$ and $\Sigma_{\mathbf{m}_t}$ are independently sampled from a Wishart distribution with an identity scale matrix and degrees of freedom equal to 5.

Selection and Perturbation Biases. Since the semantic subspace has dimension 10, the range of selection biases θ is given by $[2^{10} - 1]$, resulting in 1023 possible subsets. To systematically analyze the influence of selection bias, we choose 10 representative scenarios, detailed in Table 2.

For perturbation biases, we fix a full-selection scenario with $\theta = 1023$, which also yields 1023 possible perturbation settings. Similarly, we select 10 representative perturbation-bias scenarios, described in Table 3. Additive Gaussian noise

$$\tilde{s}_i = s_i + \mathcal{N}(0, \Sigma_\epsilon)$$

is applied independently with probability 0.75 to each semantic dimension $i \in \mathbb{I}_\rho$ in modality $\mathbf{t}$, simulating random perturbations. The noise covariance Σ_ϵ is sampled from a Wishart distribution with an identity scale matrix and degrees of freedom equal to 10.

Results from these 20 diverse settings validate our theoretical claims. To explicitly study the joint effect of selection and perturbation biases, we consider the scenario $\theta = 968$ for selection bias (excluding the last two indices) and $\rho = 12 \mid \theta = 968$ for perturbation bias (perturbing the first two indices).

Table 2: **Selection Bias Settings and Selected Semantic Indices.**

$\theta, \rho = 1$	1	11	56	176	386	638	848	968	1013	1023
$\mathbb{I}_\theta$	$\{1\}$	$[2]$	$[3]$	$[4]$	$[5]$	$[6]$	$[7]$	$[8]$	$[9]$	$[10]$

Table 3: **Perturbation Bias Settings and Perturbable Semantic Indices.**

$\rho \mid \theta = 1023$	1	2	12	57	177	387	639	849	969	1014
$\mathbb{I}_\rho$	$\emptyset$	$\{1\}$	$[2]$	$[3]$	$[4]$	$[5]$	$[6]$	$[7]$	$[8]$	$[9]$

Parameter Settings. We parameterize the generative functions $\mathbf{g}_x$ and $\mathbf{g}_{t(\theta)}$ using randomly initialized 3-layer invertible MLPs, following prior work Daunhawer et al. (2022). Invertibility is enforced by maintaining a condition number threshold of $1\text{e}{-3}$ for each layer.

The encoding functions $\mathbf{f}_x$ and $\mathbf{f}_t$ are implemented as 7-layer MLPs and optimized using the Adam optimizer. For MMCL training, we use a batch size of 6144, a learning rate of $1\text{e}{-4}$, and train for 100,000 iterations.

The loss function is given by Eq. (2), with Euclidean distance used as the similarity metric and a temperature parameter set to 1.0. To ensure training stability, gradients are clipped using a maximum 2-norm of 2.

All experiments are run with three distinct random seeds, and we report averaged R^2 values, clipped to the interval $[0, 1]$ for interpretability.

Downstream Tasks Design. To evaluate pretrained representations under various bias conditions, we construct several downstream tasks. Specifically, four regression tasks are created by generating labels using complex nonlinear functions f_{y_i}, each applied to different subsets of the true semantic variables:

$$y_1 = f_{y_1}(\mathbf{s}_{[3]}), \quad y_2 = f_{y_2}(\mathbf{s}_{[5]}), \quad y_3 = f_{y_3}(\mathbf{s}_{[7]}), \quad y_4 = f_{y_4}(\mathbf{s}_{[9]}).$$

Each function $f_{y_{(\cdot)}} : \mathbb{R}^d \to \mathbb{R}$ includes quadratic, cubic, pairwise, and triple-wise interaction terms, as well as sinusoidal, logarithmic, and exponential transformations. The full formulation for a semantic vector $\mathbf{s}_{[1:d]}$ is:

$$\begin{aligned} f_{y_{(\cdot)}}(\mathbf{s}_{[d]}) = \sum_{i=1}^{d} s_i^2 + 0.3 \sum_{i=1}^{d} s_i^3 + 0.5 \sum_{1 \le i < j \le d} s_i s_j + 0.2 \sum_{1 \le i < j < k \le d} s_i s_j s_k \\ +0.7 \sum_{i=1}^{d} (\sin(s_i) + \cos(s_i)) + 0.4 \sum_{i=1}^{d} \log(1 + |s_i|) + 0.4 \sum_{i=1}^{d} e^{-|s_i|}. \end{aligned}$$

For the classification task, labels are obtained by binarizing y_2 at its median value, which serves as the decision boundary. To simulate distribution shifts in the observations $\mathbf{x}$, we apply a skewed, heavy-tailed transformation to semantic dimensions 9 and 10:

$$s_i^{\text{ood}} = 2\,\text{sign}(s_i^{\text{id}}) \cdot |s_i^{\text{id}}|^2, \quad \text{for } i \in \{9, 10\}.$$

For both downstream tasks, we *fix* the pretrained encoders and evaluate the quality of the learned representations using a two-layer MLP as a probing model. We generate 20,480 samples as the evaluation set for training the regressors and classifiers, along with an additional 20,480 samples as the in-distribution test set. To assess OOD generalization, we generate another 20,480 samples from the shifted latent space as the OOD test set.

The regressors are trained using Mean Squared Error (MSE) loss, and the classifiers use Cross-Entropy loss, both with a learning rate of 10^{-3} and trained for 10,000 steps. We report classification performance using the average Matthews Correlation Coefficient (MCC). Importantly, in the OOD setting, we perform *no adaptation or fine-tuning*; the classifier is evaluated directly to test the generalization capability of the pretrained representations.

C.2 Additional Identification Results

Results under Perturbation Bias. As shown in Figure 7, the results of predicting true latent semantic variables under various perturbation bias settings exhibit similar trends to those observed under selection bias. Modality-specific variables are consistently discarded, unbiased semantic variables are faithfully block-identified, and misaligned semantic variables become partially identifiable in scenarios with dependent latent variables—demonstrating a consistent pattern across both image and text modalities. These findings further support and reinforce our theoretical analysis.

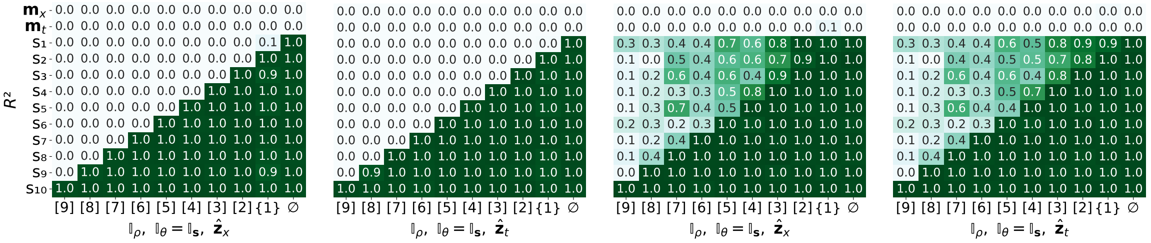


Figure 7: **Mean R^2 scores under perturbation bias settings.** Left to right: predictions using representations $\hat{\mathbf{z}}_x$ and $\hat{\mathbf{z}}_t$ under independent latent semantic variables, followed by those under dependent latent semantic variables.

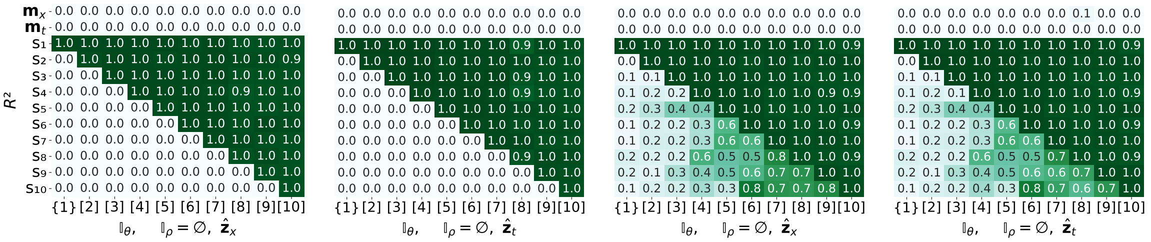


Figure 8: **Evaluating linearity of learned representations under selection bias.** Left to right: predictions using representations $\hat{\mathbf{z}}_x$ and $\hat{\mathbf{z}}_t$ under independent latent semantic variables, followed by those under dependent latent semantic variables.

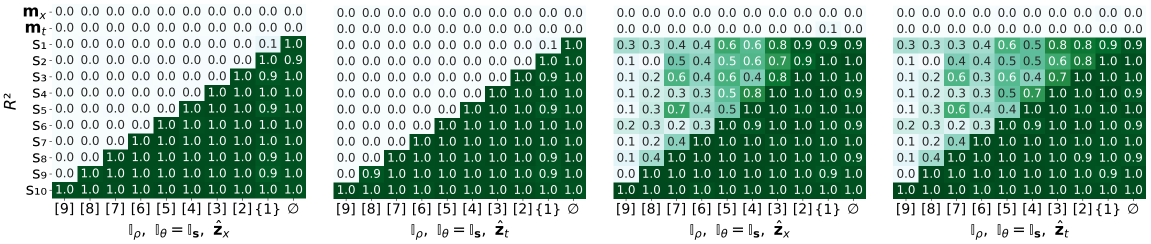


Figure 9: **Linearity of learned representations under perturbation bias.** Left to right: predictions using representations $\hat{\mathbf{z}}_x$ and $\hat{\mathbf{z}}_t$ under independent latent semantic variables, followed by those under dependent latent semantic variables.

Linearity of Learned Representations. We further analyze the linearity of the identified semantic representations by reporting the R^2 scores obtained from linear regression applied to the learned features for predicting the true latent variables. As shown in Figure 8 (selection bias) and Figure 9 (perturbation bias), the performance of linear regression closely mirrors that of nonlinear regression reported in Figure 3. This strong correspondence suggests that the relationship between the learned representations and the true latent semantic variables is approximately linear, supporting the conclusion that the identified semantic subspace is nearly linear.

Ablation Studies. We perform two ablation studies to further examine the robustness of our theoretical findings.

First, we investigate the impact of assigning an incorrect representation dimension. Specifically, we consider a scenario in which the true dimension of the selected semantic variables is 3 (with selection $\mathbb{I}_\theta = [3]$), but we intentionally set the representation dimension to 5. As shown in Table 4, in the independent case, all selected semantic variables are successfully preserved, while omitted semantic variables are effectively discarded. In contrast, the dependent scenario yields significantly different patterns of R^2 scores compared to those obtained using the correct representation dimension (see Figures 3 and 8). These results suggest that redundant representation dimensions tend to encode exogenous noise, potentially introducing unnecessary complexity into the learned representations.

Second, we explore the joint effect of selection and perturbation biases by defining a scenario with selection bias $\mathbb{I}_\theta = [8]$ and perturbation bias $\mathbb{I}_\rho = [2]$. Results presented in Table 5 demonstrate that when both biases coexist, their effects on semantic identification remain consistent: semantic variables that are either omitted or perturbed are discarded, while unbiased semantic variables—those that are both selected and unperturbed—are reliably preserved in the learned representations.

Together with previous results, these findings further reinforce our theoretical conclusions in Thm. 4.1.

C.3 Additional Downstream Results

We further report downstream task performance under varying perturbation bias settings. Specifically, the preserved semantic variables are sequentially reversed—starting from semantic index 10 and incrementally expanding until the full semantic set is included. The results shown in Figure 10 indicate that, in general,

Table 4: **The Effect of Using an Incorrect Encoding Size.** The representation size is set to 5, whereas the true dimension should be 3. The biases are defined as $\mathbb{I}_\theta = [3]$ and $\mathbb{I}_\rho = \emptyset$.

Setting		Reps.	R^2 **of Predicting Latent Semantic Variables under** $\mathbb{I}_\theta = [3]$ **and** $\mathbb{I}_\rho = \emptyset$											
			s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	$\mathbf{m}_x$	$\mathbf{m}_t$
independ.	linear	$\hat{\mathbf{z}}_x$	0.98	0.96	0.98	0.01	0.03	0.03	0.02	0.04	0.01	0.02	0.02	0.00
		$\hat{\mathbf{z}}_t$	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
	non-lin.	$\hat{\mathbf{z}}_x$	0.98	0.99	0.99	0.00	0.02	0.02	0.03	0.04	0.00	0.00	0.00	0.00
		$\hat{\mathbf{z}}_t$	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
dependent	linear	$\hat{\mathbf{z}}_x$	0.98	0.99	0.99	0.14	0.23	0.16	0.10	0.22	0.42	0.36	0.01	0.00
		$\hat{\mathbf{z}}_t$	1.00	1.00	1.00	0.13	0.20	0.13	0.09	0.20	0.41	0.34	0.00	0.01
	non-lin.	$\hat{\mathbf{z}}_x$	0.99	1.00	0.99	0.13	0.22	0.16	0.15	0.20	0.43	0.36	0.02	0.00
		$\hat{\mathbf{z}}_t$	1.00	1.00	1.00	0.13	0.20	0.13	0.09	0.20	0.42	0.35	0.00	0.05

Table 5: **Coexistence of Both Selection and Perturbation Biases.** The biases are defined as $\mathbb{I}_\theta = [8]$ and $\mathbb{I}_\rho = [2]$.

Setting		Reps.	R^2 **of Predicting Latent Semantic Variables** ($\mathbb{I}_\theta = [8], \mathbb{I}_\rho = [2]$)											
			s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	$\mathbf{m}_x$	$\mathbf{m}_t$
independ.	linear	$\hat{\mathbf{z}}_x$	0.00	0.01	0.97	0.98	0.97	0.95	0.99	0.98	0.00	0.00	0.00	0.00
		$\hat{\mathbf{z}}_t$	0.00	0.00	0.98	0.98	0.98	0.98	0.99	0.98	0.00	0.00	0.00	0.00
	non-lin.	$\hat{\mathbf{z}}_x$	0.00	0.01	0.98	0.98	0.98	0.98	0.96	0.99	0.98	0.00	0.00	0.00
		$\hat{\mathbf{z}}_t$	0.00	0.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.00	0.00	0.00
dependent	linear	$\hat{\mathbf{z}}_x$	0.67	0.57	0.97	0.99	0.99	0.98	0.99	0.97	0.63	0.63	0.00	0.00
		$\hat{\mathbf{z}}_t$	0.64	0.53	0.98	0.97	0.99	0.98	0.99	0.98	0.61	0.60	0.00	0.00
	non-lin.	$\hat{\mathbf{z}}_x$	0.68	0.57	0.99	0.99	0.99	0.99	0.99	0.99	0.64	0.64	0.00	0.00
		$\hat{\mathbf{z}}_t$	0.65	0.53	0.99	0.99	0.99	0.99	0.99	0.99	0.61	0.61	0.00	0.00

semantic variables critical to downstream tasks must be preserved in the learned representations to achieve high performance. This observation holds across both independent and dependent latent semantic settings.

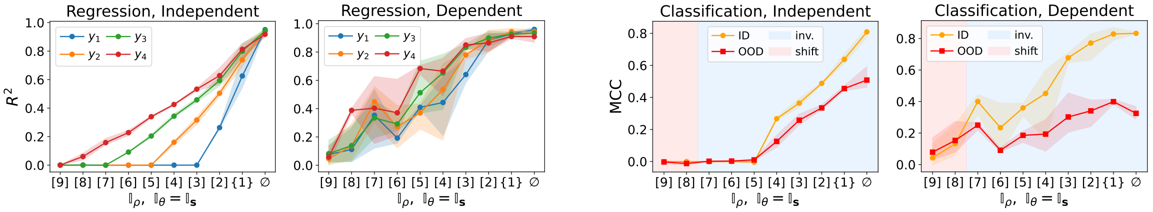


Figure 10: **Downstream performance of pretrained representations $\hat{\mathbf{z}}_x$ under perturbation bias.** Top: in-distribution (ID) regression performance. Bottom: ID classification and out-of-distribution (OOD) generalization.

D Experiment Details on MPI3D-Complex Dataset

We provide additional details on the MPI3D-Complex dataset that are not fully covered in § 5.2. In App. D.1, we comprehensively describe the experimental setup, including a dataset overview, the selection and perturbation bias configurations used to generate text, model architecture, and training parameters. In App. D.2, we present additional results, including the use of a linear classifier for predicting latent factors and an ablation study on encoder dimensionality.

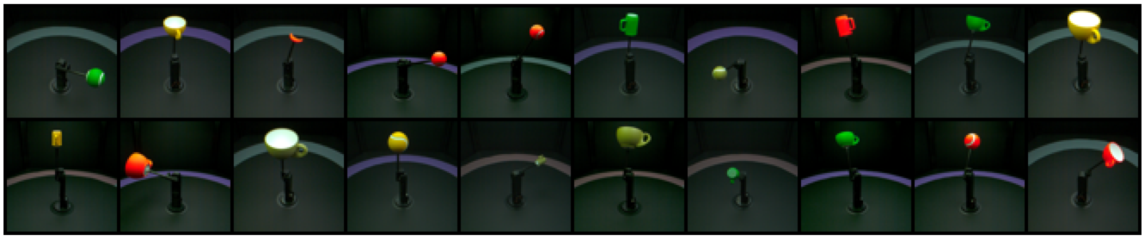


Figure 11: **Random image samples from the MPI3D-Complex dataset.** The images are captured by a camera in a controlled environment.

D.1 Detailed Experimental Setup

MPI3D-Complex Dataset. The MPI3D-Complex dataset (Gondal et al., 2019) contains 460,800 real-world images of resolution $64 \times 64 \times 3$, captured using a camera in a controlled environment. The dataset spans all possible combinations of seven *mutually independent, discrete* latent factors, as detailed in Table 6. Representative image samples are shown in Figure 11.

Due to their low-level spatial nature and the difficulty in directly mapping them to semantic descriptions, we treat the horizontal and vertical axes (`hori.`, `vert.`) as image-specific factors. The remaining factors are considered latent semantic variables for the purposes of our experiments.

Text Latent Factors. Under unbiased conditions, the text semantic variables are identical to the image semantic variables. For biased settings, we include only the content words corresponding to the selected true image semantic variables, as specified in Table 6.

To introduce linguistic variation, we employ three manually designed text templates for text generation. The choice of template is controlled by a text-specific factor m_x, sampled from a uniform distribution:

$$m_x \sim \text{Uniform}(\{1, 2, 3\}).$$

Table 6: **Latent variables of MPI3D-Complex and corresponding content words in text.**

Factor Name	Distribution	Content Words
Object color (`color`)	$\text{Uniform}(\{0, \ldots, 3\})$	`yellow`, `green`, `olive`, `red`
Object shape (`shape`)	$\text{Uniform}(\{0, \ldots, 3\})$	`coffee-cup`, `tennis-ball`, `croissant`, `beer-cup`
Object size (`size`)	$\text{Uniform}(\{0, 1\})$	`small`, `large`
Camera height (`cam.`)	$\text{Uniform}(\{0, \ldots, 2\})$	`top`, `center`, `bottom`
Background color (`back.`)	$\text{Uniform}(\{0, \ldots, 2\})$	`purple`, `sea-green`, `salmon`
Horizontal axis (`hori.`)	$\text{Uniform}(\{0, \ldots, 39\})$	— (image-specific factor)
Vertical axis (`vert.`)	$\text{Uniform}(\{0, \ldots, 39\})$	— (image-specific factor)

Text Generation under Misalignment Settings. We generate text for each image under various selection and perturbation bias settings to investigate the effects of misalignment. To ensure computational tractability—since exhaustively enumerating all possible configurations is both infeasible and unnecessary—we adopt an incremental strategy for introducing biases into the latent semantic variables, as outlined in Table 7.

In selection bias settings (where the perturbable set $\mathbb{I}_\rho$ is empty), textual descriptions include only the content words corresponding to a subset of the true image semantic variables. The specific text generation templates associated with each selection setting are detailed in Table 8.

In perturbation bias settings (where all semantic factors are selected), we introduce misalignment by randomly altering the values of the perturbable semantic variables in the text: each such value is replaced with a randomly sampled value from its support with probability 0.9.

Table 7: **Misalignment settings for text generation of MPI3D-Complex dataset.**

Setting	Selection Bias, $\mathbb{I}_\theta$	Perturbation Bias, $\mathbb{I}_\rho$
①	{color}	$\emptyset$
②	{color, shape}	{back.}
③	{color, shape, size}	{cam., back.}
④	{color, shape, size, cam.}	{size, cam., back.}
⑤	{color, shape, size, cam., back.}	{shape, size, cam., back.}

Table 8: **Text generation templates for MPI3D-Complex dataset for different selection settings.**

Setting	Text Generation Templates for Each Selection View
①	"An object colored {color}." "It has a {color} appearance." "Something with {color}."
②	"A {shape} that is {color}." "The {color} {shape}." "An object shaped like a {shape}, colored {color}."
③	"A { size} {shape} in {color}." "{color}, { size}, {shape}." "The object is { size}, shaped as a {shape}, and colored {color}."
④	"A { size}{shape} in {color}, seen from {cam.}." "Viewed from {cam.}, a {color}, { size} {shape}." "A { size} {shape} with {color}, perspective: {cam.}."
⑤	"A { size} {shape} in {color}, viewed from {cam.}, {back.}." "From {cam.}, you see a {color}, { size} {shape}, {back.}." "A { size} {shape}, {color}, placed {back.}, observed from {cam.}."

Training Details. For each setting, the dataset is partitioned into training, evaluation, and test subsets in a fixed ratio of $44{,}720 : 23{,}040 : 23{,}040$. Across all configurations, we train the image and text encoders for $200{,}000$ steps using the training subset, averaging results over three random seeds. The training objective is the multimodal contrastive loss $\mathcal{L}_{\text{MMCL}}$, defined in Eq. (1), and optimized using the Adam optimizer with an initial learning rate of 1×10^{-5}, a batch size of 256, and a temperature parameter $\tau = 1.0$.

For image encoding, we use a ResNet18 backbone followed by a fully connected layer with a fixed input dimensionality of 100. The output dimensionality is adjusted according to the number of unbiased semantic factors under each bias setting.

For text encoding, we tokenize text using the `nltk.PunktTokenizer`, following the procedure in Daunhawer et al. (2022). Tokenized sequences are transformed into two-dimensional one-hot embeddings and processed using a convolutional neural network (CNN) with a variable number of layers, determined by the shape of the tokenized input. The output dimensionality of the CNN is configured to match that of the image encoder, ensuring compatibility in the shared representation space.

The encoding size is generally set to match the number of unbiased semantic dimensions, i.e., $\dim(\mathbb{I}_\rho^c)$. However, when this number is less than 3, we set the encoding size to 3 to ensure minimal representational capacity. An ablation study on this design choice is provided in the following section.

Evaluation Metrics. After training the representations, we freeze the encoders and train both a linear classifier (logistic regression) and a nonlinear classifier (a two-layer MLP with ReLU activation) for each setting and each image latent factor. Classifiers are trained on the evaluation subset for $10{,}000$ steps.

Performance is assessed on the test set using the Matthews Correlation Coefficient (MCC), computed separately for each latent factor. MCC is chosen for its robustness in evaluating binary classification performance under class imbalance.

D.2 Additional Results

Linearity of Learned Representations. We evaluate the linear separability of the learned representations by reporting MCC scores for predicting each latent factor using a linear classifier. As shown in Figure 12, and in comparison to the nonlinear results in Figure 5, the findings reveal that certain latent semantic variables—most notably `size`—are not linearly embedded in the learned representation space when training image-text pairs with MMCL.

In contrast, factors such as `color`, `shape`, `cam.`, and `back.` exhibit strong linear separability, suggesting that these semantic variables are linearly represented by the learned image and text encoders.

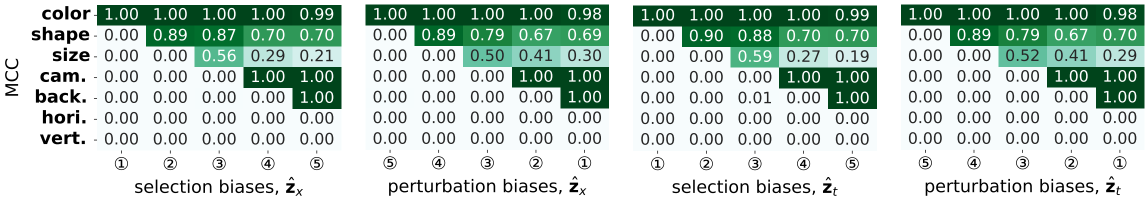


Figure 12: **Evaluating linearity of learned representations under misalignment.** Left to right: image features $\hat{\mathbf{z}}_x$ under selection bias, image features $\hat{\mathbf{z}}_x$ under perturbation bias, text features $\hat{\mathbf{z}}_t$ under selection bias, and text features $\hat{\mathbf{z}}_t$ under perturbation bias.

Ablations on the Encoding Size. We conduct an ablation study on the encoding size by evaluating a selection bias setting in which only the semantic attribute {`color`} is selected for text generation. All other training parameters are kept consistent with those used in our main experiments, except for the encoding dimensionality.

As shown in Figure 13, in contrast to training with purely numerical data, learning from image-text data exhibits sensitivity to the choice of encoding size. Specifically, when the encoding size is set to 1—exactly matching the number of perfectly aligned semantic dimensions—the image encoder tends to be under-optimized, resulting in high variance across runs. However, increasing the encoding size leads to more stable and reliable performance.

Notably, in the presence of independent latent factors, the additional (redundant) encoding dimensions do not appear to capture misaligned semantic variables, suggesting that excess capacity does not harm identifiability in this setting.

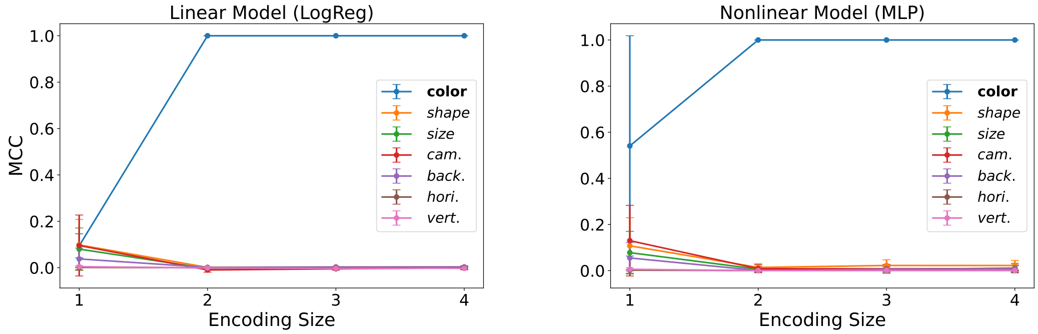


Figure 13: **Ablation study on encoding size** under a selection bias setting where only the {`color`} attribute is included in text generation. Left: average MCC over three runs using a linear classifier. Right: average MCC using a nonlinear classifier.

E Experiment Details on Causal3DIdent Dataset

We provide additional details on the Causal3DIdent dataset that are not fully covered in § 5.3. In App. E.1, we offer a comprehensive description of the experimental setup, including the image and text latent factors, the image generation process, the design of selection and perturbation bias settings for text generation, and training configurations. In App. E.2, we present supplementary results, including analyses of the learned text representations and assessments of the linearity of the learned embeddings.

E.1 Detailed Experimental Setup

Image Latent Factors and Image Generation. Following prior work Zimmermann et al. (2021); Von Kügelgen et al. (2021); Daunhawer et al. (2022); Yao et al. (2023), we utilize the *Causal3DIdent* dataset to synthesize images from a predefined latent causal structure. Images are generated using the Blender renderer Community (2018), which applies a complex rendering function parameterized by 11 input variables. In our configuration, the object's z-position is fixed, leaving 10 latent factors that govern image generation.

These include 3 discrete variables—object shape (`shape`), and object positions along the horizontal (`x_pos`) and vertical (`y_pos`) axes—and 7 continuous variables: object color (`color`), spotlight position (`s_pos`) and color (`s_color`), background color (`b_color`), and the three object rotation angles (`alpha`, `beta`, `gamma`). We treat the rotation angles (`alpha`, `beta`, `gamma`) as image-specific latent variables, while the remaining factors are considered semantic latent variables, structured according to the causal graph shown in Figure 14.

We synthesize 80,000 samples for MMCL training, 10,000 samples for classifier or regressor training, and another 10,000 samples for test-time evaluation. Images are rendered at a resolution of $128 \times 128 \times 3$.

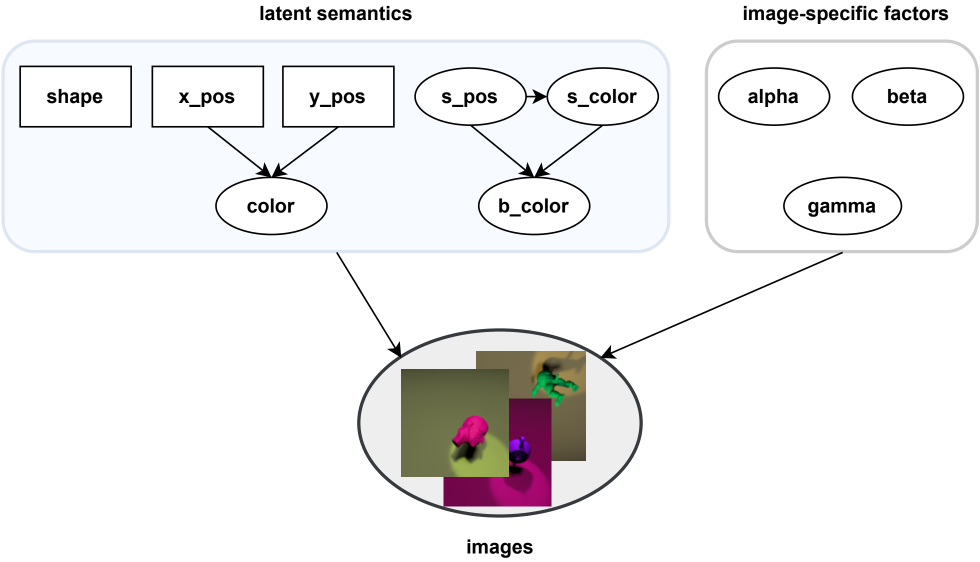


Figure 14: **Latent causal model governing image generation in the Causal3DIdent dataset.** Rectangular nodes represent discrete latent random variables, while elliptical nodes denote continuous ones. Object shape (`shape`), horizontal and vertical position (`x_pos`, `y_pos`), object color (`color`), spotlight position and color (`s_pos`, `s_color`), and background color (`b_color`) are treated as latent semantic variables shared across modalities and potentially subject to misalignment. In contrast, the rotation angles—`alpha`, `beta`, and `gamma`—are considered image-specific latent factors.

Text Latent Factors. We discretize the continuous variables `color`, `s_color`, and `b_color` using sampled image semantic variables mapped to distinct color palettes: `TABLEAU_COLORS` for `color`, `CSS4_COLORS` for `s_color`, and `XKCD_COLORS` for `b_color`. While `s_color` remains continuous in the underlying latent representation, we simulate partial information loss for the spotlight position (`s_pos`) during the mapping to text. As a result, the generated textual descriptions of `s_pos` do not constitute an information-preserving transformation.

To introduce text-specific variation, we employ five manually designed templates to generate text from the latent factors under each bias setting, adapting the text rendering pipeline from (Daunhawer et al., 2022). A complete list of latent factors and their types is provided in Table 9.

Text Generation under Different Misalignment Settings. Following the MPI3D-Complex experiments, we explore a series of incrementally increasing selection and perturbation bias configurations to introduce varying degrees and types of cross-modal misalignment. These settings enable a systematic investigation of

Table 9: **Latent factors of Causal3DIdent and corresponding content words in text.**

Factor Name	Image Modality	Text Modality	Content Words
`shape`	$\text{Uniform}(\{0,\ldots,6\})$	$\text{Uniform}(\{0,\ldots,6\})$	`teapot`, `hare`, `dragon`, `cow`, `armadillo`, `horse`, `head`
`x_pos`	$\text{Uniform}(\{0,1,2\})$	$\text{Uniform}(\{0,1,2\})$	`left`, `center`, `right`
`y_pos`	$\text{Uniform}(\{0,1,2\})$	$\text{Uniform}(\{0,1,2\})$	`top`, `mid`, `bottom`
`s_pos`	$\text{Uniform}([0,1])$	$\text{Uniform}([0,1])$	`northwest`, `northeast`, `center`, `southwest`, `southeast`
`color`	$\frac{1}{6}(\texttt{x\_pos}+\texttt{y\_pos})+\frac{1}{3}\text{Uniform}([0,1])$	Up to 10 colors	Color names in `TABLEAU_COLORS`
`s_color`	$\frac{1}{2}(\texttt{s\_pos}+\text{Uniform}([0,1]))$	Up to 147 colors	Color names in `CSS4_COLORS`
`b_color`	$\frac{1}{3}(\texttt{s\_pos}+\texttt{s\_color}+\text{Uniform}([0,1]))$	Up to 954 colors	Color names in `XKCD_COLORS`
`alpha`	$\text{Uniform}([0,1]))$	—	—
`beta`	$\text{Uniform}([0,1]))$	—	—
`gamma`	$\text{Uniform}([0,1]))$	—	—
`phrase`	—	$\text{Uniform}(\{0,\ldots,5\}))$	—

how different forms of alignment impact representation learning. Each configuration is indexed using circled numerals and summarized in Table 10.

For the perturbable semantic variables in each perturbation setting, we randomly sample the corresponding text semantic values. Specifically, for discrete variables such as `x_pos` and `y_pos`, we sample uniformly from the set $\{0, 1, 2\}$; for continuous variables, we sample uniformly from the interval $[0, 1]$.

Table 11 provides the text generation templates associated with each selection setting. Representative image–text pairs generated under different selection and perturbation bias configurations are shown in Figure 15.

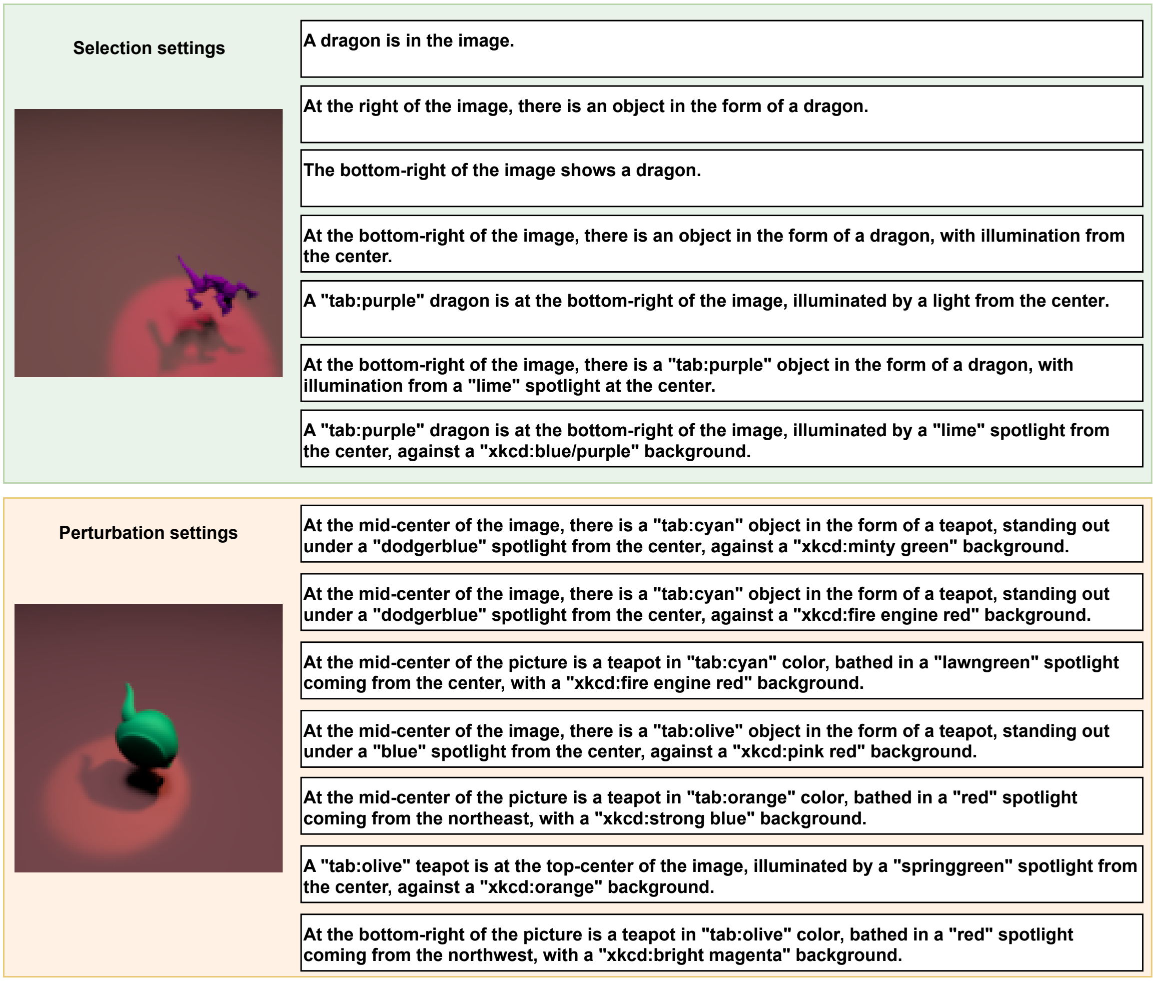
Table 10: **Misalignment settings for text generation of Causal3DIdent dataset.**

Setting	Selection Bias, $\mathbb{I}_\theta$	Perturbation Bias, $\mathbb{I}_\rho$
①	{`shape`}	{`x_pos`, `y_pos`, `s_pos`, `color`, `s_color`, `b_color`}
②	{`shape`, `x_pos`}	{`y_pos`, `s_pos`, `color`, `s_color`, `b_color`}
③	{`shape`, `x_pos`, `y_pos`}	{`s_pos`, `color`, `s_color`, `b_color`}
④	{`shape`, `x_pos`, `y_pos`, `s_pos`}	{`color`, `s_color`, `b_color`}
⑤	{`shape`, `x_pos`, `y_pos`, `s_pos`, `color`}	{`s_color`, `b_color`}
⑥	{`shape`, `x_pos`, `y_pos`, `s_pos`, `color`, `s_color`}	{`b_color`}
⑦	{`shape`, `x_pos`, `y_pos`, `s_pos`, `color`, `s_color`, `b_color`}	$\emptyset$

Training Details. Across all experimental settings, we train the image and text encoders for 100,000 steps on the training subset, using three different random seeds to ensure robustness. The training objective is the multimodal contrastive loss $\mathcal{L}_{\text{MMCL}}$, defined in Eq. (1), and optimized using the Adam optimizer with an initial learning rate of 1×10^{-5}, a batch size of 256, and a temperature parameter $\tau = 1.0$.

For both the image and text encoders, we adopt the same architectures used in the MPI3D-Complex experiments. The encoding dimensionality is adjusted according to the bias setting: for selection settings ① through ⑦, the encoding sizes are set to 3, 3, 4, 5, 5, 6, and 7, respectively; for perturbation settings ① through ⑦, the encoding sizes are assigned in reverse order: 7, 6, 5, 5, 4, 3, and 3.

Evaluation Metrics. After training the representations, we freeze the encoders and, for each bias setting, train both a linear classifier (logistic regression) and a nonlinear classifier (a two-layer MLP with ReLU activation) for each discrete latent factor. Similarly, for continuous latent factors, we train both a linear regressor and a nonlinear regressor (a two-layer MLP with ReLU activation). All classifiers and regressors are trained for 10,000 steps using the evaluation subset.



A dragon is in the image.

At the right of the image, there is an object in the form of a dragon.

The bottom-right of the image shows a dragon.

At the bottom-right of the image, there is an object in the form of a dragon, with illumination from the center.

A "tab:purple" dragon is at the bottom-right of the image, illuminated by a light from the center.

At the bottom-right of the image, there is a "tab:purple" object in the form of a dragon, with illumination from a "lime" spotlight at the center.

A "tab:purple" dragon is at the bottom-right of the image, illuminated by a "lime" spotlight from the center, against a "xkcd:blue/purple" background.

At the mid-center of the image, there is a "tab:cyan" object in the form of a teapot, standing out under a "dodgerblue" spotlight from the center, against a "xkcd:minty green" background.

At the mid-center of the image, there is a "tab:cyan" object in the form of a teapot, standing out under a "dodgerblue" spotlight from the center, against a "xkcd:fire engine red" background.

At the mid-center of the picture is a teapot in "tab:cyan" color, bathed in a "lawngreen" spotlight coming from the center, with a "xkcd:fire engine red" background.

At the mid-center of the image, there is a "tab:olive" object in the form of a teapot, standing out under a "blue" spotlight from the center, against a "xkcd:pink red" background.

At the mid-center of the picture is a teapot in "tab:orange" color, bathed in a "red" spotlight coming from the northeast, with a "xkcd:strong blue" background.

A "tab:olive" teapot is at the top-center of the image, illuminated by a "springgreen" spotlight from the center, against a "xkcd:orange" background.

At the bottom-right of the picture is a teapot in "tab:olive" color, bathed in a "red" spotlight coming from the northwest, with a "xkcd:bright magenta" background.

Figure 15: **Example image-text pairs from Causal3DIdent under different selection bias settings.** The left panel shows randomly selected images; the right panel presents the corresponding text from top to bottom, each generated under selection settings ① to ⑦ with no perturbations, and perturbation bias ⑦ to ① with full selections.

We assess the predictive performance of the learned representations by evaluating their ability to recover the ground-truth latent factors corresponding to their respective modalities. This evaluation accounts for the fact that some semantic variables may appear in discrete or continuous form, depending on the modality and rendering process.

Prediction performance is measured on the test set using the Matthews Correlation Coefficient (MCC) for discrete factors and the coefficient of determination (R^2) for continuous factors. Metrics are computed separately for each latent factor.

E.2 Additional Results

Results of Text Representations. We now turn to the analysis of the text representations learned by MMCL. As shown in Figure 16, we observe patterns similar to those found in the image modality. In particular, discrete latent semantic variables—such as `shape`, `x_pos`, and `y_pos`—are reliably identified, provided they are unbiased and consistently aligned across modalities. A notable case is `s_pos`, which is a continuous latent factor in both modalities but is mapped to only five discrete tokens in the text observations, rendering the text generation process non-invertible. Despite this lossy transformation, the model achieves a relatively high R^2, suggesting that the learned text representations remain strongly influenced by the alignment objective, even when semantic information is partially lost.

Table 11: **Text generation templates for Causal3DIdent dataset under different selection settings.**

Setting	Text Generation Templates
①	"A {`shape`} is visible." "A {`shape`} is in the image." "The image shows a {`shape`}." "The picture is a {`shape`}." "There is an object in the form of a {`shape`}."
②	"A {`shape`} is visible, positioned at the {`x_pos`} of the image." "A {`shape`} is at the {`x_pos`} of the image." "The {`x_pos`} of the image shows a {`shape`}." "At the {`x_pos`} of the picture is a {`shape`}." "At the {`x_pos`} of the image, there is an object in the form of a {`shape`}."
③	"A {`shape`} is visible, positioned at the {`y_pos`}-{`x_pos`} of the image." "A {`shape`} is at the {`y_pos`}-{`x_pos`} of the image." "The {`y_pos`}-{`x_pos`} of the image shows a {`shape`}." "At the {`y_pos`}-{`x_pos`} of the picture is a {`shape`}." "At the {`y_pos`}-{`x_pos`} of the image, there is an object in the form of a {`shape`}."
④	"A {`shape`} is visible, positioned at the {`y_pos`}-{`x_pos`}, with a spotlight shining from {`s_pos`}." "A {`shape`} is at the {`y_pos`}-{`x_pos`}, illuminated by a light from {`s_pos`}." "The {`y_pos`}-{`x_pos`} shows a {`shape`}, highlighted by a light from {`s_pos`}." "At the {`y_pos`}-{`x_pos`} is a {`shape`}, under a light from {`s_pos`}." "There is a {`shape`} at {`y_pos`}-{`x_pos`}, lit from {`s_pos`}."
⑤	"A {`shape`} of {`color`} color is visible at {`y_pos`}-{`x_pos`}, with a spotlight from {`s_pos`}." "A {`color`} {`shape`} is at {`y_pos`}-{`x_pos`}, lit from {`s_pos`}." "The area {`y_pos`}-{`x_pos`} shows a {`color`} {`shape`}, under a light from {`s_pos`}." "A {`color`} {`shape`} is illuminated at {`y_pos`}-{`x_pos`} from {`s_pos`}." "A {`color`} object shaped like a {`shape`} is lit from {`s_pos`}."
⑥	"A {`color`} {`shape`} is lit by a {`s_color`} spotlight from {`s_pos`}, at {`y_pos`}-{`x_pos`}." "At {`y_pos`}-{`x_pos`}, a {`color`} {`shape`} is under a {`s_color`} light from {`s_pos`}." "The {`shape`} is {`color`}, under a {`s_color`} light at {`s_pos`}." "A {`color`} {`shape`} under a {`s_color`} spotlight at {`s_pos`}, located at {`y_pos`}-{`x_pos`}." "A {`color`} {`shape`} stands under a {`s_color`} light from {`s_pos`}."
⑦	"A {`color`} {`shape`} under a {`s_color`} spotlight at {`s_pos`}, with a {`b_color`} background, at {`y_pos`}-{`x_pos`}." "At {`y_pos`}-{`x_pos`}, a {`color`} {`shape`} is under a {`s_color`} light from {`s_pos`}, against a {`b_color`} background." "A {`color`} {`shape`} appears at {`y_pos`}-{`x_pos`}, lit by {`s_color`} from {`s_pos`}, with {`b_color`} background." "The scene shows a {`color`} {`shape`} under {`s_color`} lighting at {`s_pos`}, with a {`b_color`} backdrop." "A {`color`} object shaped like a {`shape`}, under a {`s_color`} spotlight at {`s_pos`}, with a {`b_color`} background."

For other latent semantic variables that are continuous in the image modality but discretized in the text modality—such as `color`, `s_color`, and `b_color`—we observe varying degrees of performance degradation. This drop in performance is likely attributable not only to the quantization of continuous values but also to semantic ambiguity introduced during text generation. Notably, all three attributes correspond to different aspects of color, yet they may be described using overlapping vocabulary drawn from distinct color palettes. For instance, `tab:cyan` from `TABLEAU_COLORS` refers to object color (`color`), `cyan` from `CSS4_COLORS` describes spotlight color (`s_color`), and `xkcd:cyan` from `XKCD_COLORS` indicates background color (`b_color`). Despite referencing different latent variables, these tokens all contain the word `cyan`, which is tokenized identically by `nltk.PunktTokenizer`, resulting in ambiguity in the text observations. These findings highlight the importance of using distinct and unambiguous content words when representing semantically different concepts in multimodal learning—particularly when the text modality is not treated merely as an auxiliary input for visual representation learning.

Interestingly, the identification of x_pos and y_pos in the text representations does not lead to improved predictability of color, in contrast to what is often observed in the image modality. This aligns with our theoretical expectation that perturbation biases disrupt the underlying causal structure in the image latents.

Regarding the text-specific factor phrase, we find it to be partially encoded in the learned representations. This contrasts with the image-specific continuous factors, which are consistently omitted. The partial identifiability of phrase is likely attributable to its discrete nature, which violates the conditions typically required for modality-specific factors to be excluded—consistent with findings reported in Daunhawer et al. (2022). Moreover, this effect appears more pronounced under selection bias settings, particularly when the encoding dimensionality exceeds the true dimensionality of the unbiased semantic subspace. Overall, the behavior of the learned text representations provides empirical support for our theoretical analysis, even under conditions where certain modeling assumptions are relaxed or violated.

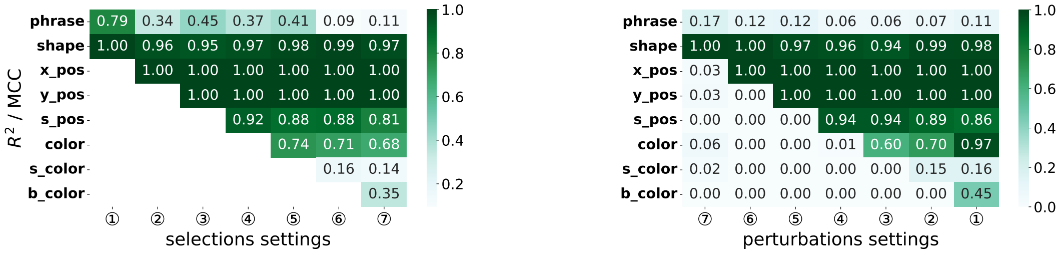


Figure 16: **Prediction of latent semantic variables under misalignment with text features.** R^2 for continuous and MCC for discrete factors. Left: Selection bias. Right: Perturbation bias.

Linearity of Learned Representations. We assess the linear separability of the learned image and text representations by evaluating the performance of linear classifiers and regressors trained to predict each latent factor. As shown in Figure 17 and Figure 18, and in comparison to the nonlinear prediction results, the findings indicate that not all latent semantic variables are linearly embedded in the representations learned through MMCL. In particular, the semantic factor color consistently demonstrates poor linear predictability, suggesting that it is encoded nonlinearly in both modalities. Conversely, factors such as x_pos and y_pos exhibit strong linear separability in certain settings, indicating that these semantic variables are more directly captured in the latent space of both the image and text encoders. Overall, these results suggest that the linearity of the learned representations is factor-dependent and shaped by both modality-specific encoding strategies and the underlying structure of the input data—highlighting an important direction for future investigation.

F Computation Resources

All experiments were conducted on a high-performance computing cluster equipped with 4×NVIDIA A100 GPUs (40 GB each), running CUDA 12.2 and driver version 535.161.07. The system also included an AMD EPYC 7313 16-core processor and 503 GB of RAM. For the numerical simulations, we trained over 120 models in total, requiring approximately 70 GPU-hours across 4 GPUs. On the MPI3D-Complex dataset, we trained 36 models, consuming approximately 27 GPU-hours. For the Causal3DIdent dataset, we trained 42 models, which required roughly 25 GPU-hours across 4 GPUs. Additionally, we generated 100,000 synthetic images for the Causal3DIdent dataset using Blender. Rendering was performed over four days on a separate workstation equipped with an AMD Ryzen 7 7700X 8-core processor (4.50 GHz) and a single NVIDIA RTX 4090 GPU (24 GB).

G Discussion: Limitations and Future Directions

In this section, we reflect on the limitations of the current study, propose future research directions informed by our findings, and discuss broader implications.

G.1 Modeling Randomly Missing Latent Semantic Variables

Our current framework primarily addresses misalignment characterized by fixed selection and perturbation biases in text generation. However, large-scale, real-world datasets—particularly user-generated corpora

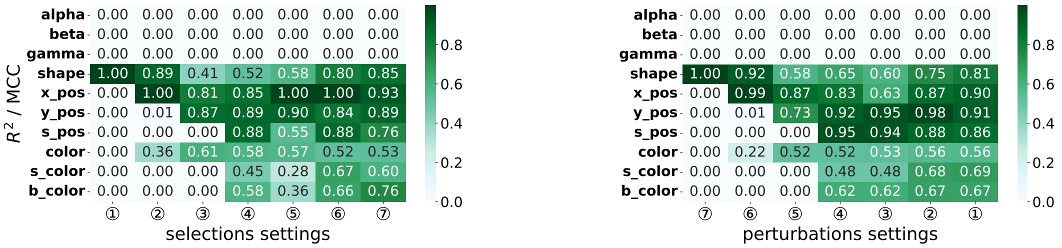


Figure 17: **Evaluating linearity of image features under misalignment settings.** R^2 for continuous and MCC for discrete factors. Left: Selection bias. Right: Perturbation bias.

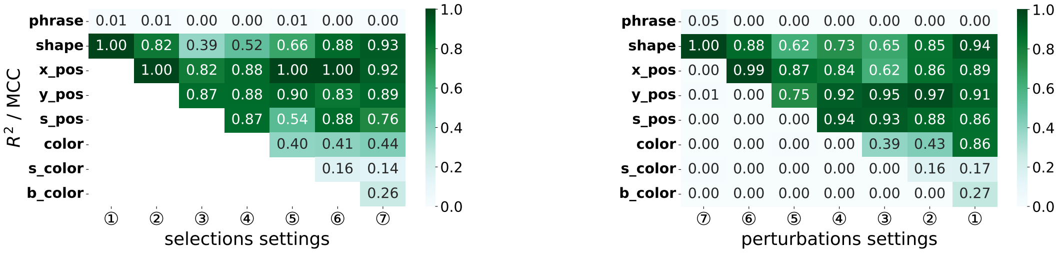


Figure 18: **Evaluating linearity of text features under misalignment settings.** R^2 for continuous and MCC for discrete factors. Left: Selection bias. Right: Perturbation bias.

such as LAION-5B Schuhmann et al. (2022)—often exhibit random and unstructured semantic omissions. Extending our latent variable model to account for such missingness (e.g., missing completely at random or missing not at random) introduces both theoretical and practical challenges. Future research should investigate the identifiability implications of randomly missing semantic variables and determine conditions under which partial or probabilistic recovery of latent factors remains tractable.

G.2 Toward the Linearity of Multimodal Representations

While our theoretical results establish identifiability of unbiased semantic factors up to invertible transformations, empirical evidence suggests that learned representations are often approximately linear with respect to the underlying semantic variables. This observation raises an important open question: under what conditions can identifiability up to a *linear* transformation be guaranteed, and can such guarantees be derived without imposing overly restrictive assumptions on the data-generating process or latent structure? Pursuing this question is both theoretically compelling and practically significant. From a practical standpoint, linear representations enhance interpretability and facilitate the use of simpler downstream models. Future work could develop training objectives, regularization schemes, or architectural inductive biases that explicitly favor linearity in learned representations. Bridging this empirical regularity with theoretical guarantees represents a promising and impactful research direction.

G.3 Linguistic Relativity and Epistemic Constraints in Multimodal AI

Our findings carry broader implications, offering computational evidence in support of a form of linguistic relativity (Whorf, 2012). Whorf posited that language shapes human perception and conceptualization of reality. Analogously, linguistic supervision in multimodal AI systems governs which aspects of the visual world are encoded—and which are systematically excluded. In this view, selection and perturbation biases in textual annotations function as *epistemic filters* that guide machine perception and understanding. This reframes dataset design as an epistemic endeavor: the inclusion or omission of content in linguistic annotations implicitly defines the conceptual space that models can represent and generalize over. Two actionable insights emerge from this perspective. First, achieving faithful and generalizable AI representations requires that linguistic supervision explicitly and comprehensively encode the semantic concepts intended for capture. Second, curating image captions to intentionally filter out harmful or undesirable content offers a mechanism for shaping representations toward ethically and socially responsible outcomes. Designing datasets guided by these dual objectives—semantic fidelity and ethical alignment—can foster the development of multimodal AI systems that both reflect human understanding and uphold human values.