MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Aligning Multimodal Representations through an Information Bottleneck

Almudévar, Antonio; Hernández-Lobato, José, M; Khurana, Sameer; Marxer, Ricard; Ortega, Alfonso

TR2025-109 July 16, 2025

Abstract

Contrastive losses have been extensively used as a tool for multimodal representation learning. However, it has been empirically observed that their use is not effective to learn an aligned representation space. In this paper, we argue that this phenomenon is caused by the presence of modality-specific information in the representation space. Although some of the most widely used contrastive losses maximize the mutual in- formation between representations of both modalities, they are not designed to remove the modality- specific information. We give a theoretical description of this problem through the lens of the Information Bottleneck Principle. We also empirically analyze how different hyperparameters affect the emergence of this phenomenon in a con- trolled experimental setup. Finally, we propose a regularization term in the loss function that is de- rived by means of a variational approximation and aims to increase the representational alignment. We analyze in a set of controlled experiments and real-world applications the advantages of includ- ing this regularization term.

International Conference on Machine Learning (ICML) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Aligning Multimodal Representations through an Information Bottleneck

Antonio Almudévar [1] **José Miguel Hernández-Lobato** [2] **Sameer Khurana** [3] **Ricard Marxer** [4] **Alfonso Ortega** [1]

Abstract

Contrastive losses have been extensively used as a tool for multimodal representation learning. However, it has been empirically observed that their use is not effective to learn an aligned representation space. In this paper, we argue that this phenomenon is caused by the presence of modality-specific information in the representation space. Although some of the most widely used contrastive losses maximize the mutual information between representations of both modalities, they are not designed to remove the modality-specific information. We give a theoretical description of this problem through the lens of the Information Bottleneck Principle. We also empirically analyze how different hyperparameters affect the emergence of this phenomenon in a controlled experimental setup. Finally, we propose a regularization term in the loss function that is derived by means of a variational approximation and aims to increase the representational alignment. We analyze in a set of controlled experiments and real-world applications the advantages of including this regularization term.

1. Introduction

Multimodal Learning is an area of AI that is focused on processing and integrating information from multiple modalities (e.g., text, image or audio). It is becoming a pivotal topic in the community because of multiple reasons, including, but not limited to, (i) it permits to mimic human cognition processes (Fei et al., 2022; Lee et al., 2023); (ii) it allows to use a greater amount of training data from different modalities, which tends to improve the performance of the models (Kaplan et al., 2020; Cuervo & Marxer, 2024); and (iii) it is essential in real-world applications such as autonomous vehicles (Xiao et al., 2020), healthcare (Kline et al., 2022) or human-computer interaction (Sinha et al., 2010).

Similarly to humans, most AI systems work through obtaining intermediate representations, which are compressed versions of the raw data that preserve useful information to solve different downstream tasks (Bengio et al., 2013; Cadieu et al., 2014). One of the most widely used ways of training multimodal systems is Contrastive Representation Learning (Karpathy & Fei-Fei, 2015; Oord et al., 2018; Tian et al., 2020a). In this paradigm, representations corresponding to similar input data are brought closer than dissimilar ones. For example, the caption "a photo of a dog" should become closer to an image of *a dog* than to that of *a cat*. The most widely used of the contrastive losses is the InfoNCE (Oord et al., 2018). Minimizing this loss is equivalent to maximizing a lower bound of the mutual information of the representations from both modalities. In other words, when minimizing this loss, representations from each modality should maximize the information that they contain about what is common between them.

However, the above does not imply that representations from both modalities contain the same information. Representations could contain all the information about what is common to both modalities, but still preserve much of the information that is specific to their own modality (a.k.a. nuisances from now on). We argue that this can translate into a substantial representational misalignment (Klabunde et al., 2023), especially when the inputs contain a high level of nuisances. In other words, representations from two modalities of a positive pair could be not so similar to each other due to the fact that they are encoding different information. Figure 1 illustrates a trivial example in which two similar, yet different, images have exactly the same caption. Thus, the text representations are exactly the same, while the image representations are different from each other, since they can be encoding information about aspects like the color of the dog, the number of clouds in the sky or the number of blades of grass. This misalignment phenomenon has been already observed and denominated *modality gap* (Liang et al., 2022). However, to the best of our knowledge, the present is the first work in which this phenomenon is explained from an information theory perspective.

[1]ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain [2]University of Cambridge, Cambridge, UK [3]Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA [4]Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon, France. Correspondence to: Antonio Almudévar <almudevar@unizar.es>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

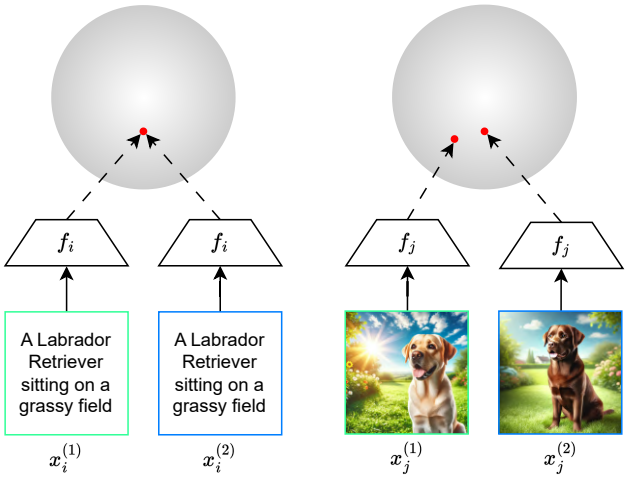


Figure 1: Different modalities usually contain different information. A trivial example of this is the case in which two different images have exactly the same caption. As a consequence of this, representations from different modalities tend to contain different information (thus leading to misalignment) if the opposite is not explicitly imposed.

It is precisely this explanation which allows us to propose a solution to this phenomenon. Concretely, we propose to apply an Information Bottleneck (IB) (Tishby et al., 2000) in the representation space. With this, apart from maximizing the mutual information between the representations of both modalities through the contrastive loss, we reduce the nuisances that can be found in the representations. This IB is applied by means of a regularization term in the loss function that is derived using a variational approximation. Thus, it is efficient, straightforward to implement and modality-agnostic, which is advantageous over alternative approaches (Li et al., 2021).

2. Preliminaries

Contrastive Representation Learning (CRL) This encompasses a set of techniques that learns a representation space in which representations of similar inputs are closer than those of dissimilar ones. It has emerged as one of the most competitive methods for learning representations without labels in a self-supervised way (Oord et al., 2018; Hjelm et al., 2018; Wu et al., 2018; Logeswaran & Lee, 2018; Bachman et al., 2019; Tian et al., 2020a; Chen et al., 2020a; Henaff, 2020). The most widely used among the contrastive losses is the InfoNCE (Oord et al., 2018) and it has been shown that minimizing this is equivalent to maximizing a lower bound of the mutual information (MI) between a pair of representations (Bachman et al., 2019; Tian et al., 2020a).

Multimodal Contrastive Representation Learning One of the tasks in which contrastive losses have gained popularity is Multimodal Representation Learning, which consists in designing systems that map inputs from different modalities (e.g. image and text) into a joint representation space. Some foundation works used rank-based losses (Yager, 1988; Usunier et al., 2009; Schroff et al., 2015) to learn multimodal representation spaces (Weston et al., 2010; Frome et al., 2013; Karpathy & Fei-Fei, 2015) while more modern approaches have used the InfoNCE loss (Tian et al., 2020a; Radford et al., 2021; Jia et al., 2021; Xu et al., 2021; Girdhar et al., 2023) for this purpose. However, it has been observed that, when trained in a contrastive way, representations from different modalities tend to be located in different regions of the space, a phenomenon called *modality gap* (Liang et al., 2022; Udandarao, 2022; Ramasinghe et al., 2024; Fahim et al., 2024; Schrodi et al., 2024). This phenomenon can be an issue in some applications such as Image Captioning or Visual Question Answering, so sophisticated training methods have been proposed to palliate it (Chen et al., 2020b; Li et al., 2021; 2022; 2023).

Measuring Representational Alignment The representational alignment (or similarity) is typically measured through kernel alignment metrics (Cristianini et al., 2001; Cortes et al., 2012). Examples of these include Centered Kernel Alignment (CKA) (Kornblith et al., 2019), SVCCA (Raghu et al., 2017) and nearest-neighbor metrics (Klabunde et al., 2023). However, in this work we restrict our attention to the former, since it is the most widely used for this purpose. Let $Z^{(\alpha)} \in \mathbb{R}^{n \times d_\alpha}$ and $Z^{(\beta)} \in \mathbb{R}^{n \times d_\beta}$ be two sets of representations, $K = k(z_i^{(\alpha)}, z_j^{(\alpha)})$ and $L = l(z_i^{(\beta)}, z_j^{(\beta)})$, where $k : \mathbb{R}^{d_\alpha} \times \mathbb{R}^{d_\alpha} \to \mathbb{R}$ and $l : \mathbb{R}^{d_\beta} \times \mathbb{R}^{d_\beta} \to \mathbb{R}$ are kernel functions. Then, the Hilbert-Schmidt Independence Criterion is defined as:

$$HSIC(K, L) = \frac{1}{(n-1)^2} \operatorname{Tr}(KHLH) \tag{1}$$

where $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. Then, the CKA is defined as follows:

$$CKA(K, L) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K)HSIC(L, L)}} \tag{2}$$

This metric ranges between zero and one and we say that a pair of representations is perfectly aligned when $CKA = 1$.

Information Bottleneck It is a framework that aims to find a representation that contains all the information in an input that is necessary to solve a given task, while discarding irrelevant information (Tishby et al., 2000; Tishby & Zaslavsky, 2015). Given an input X, a task Y and a representation Z, the goal can be formulated as:

$$\max_Z I(Z; Y) - \beta I(Z; X) \tag{3}$$

where β is a Lagrange multiplier that controls the trade-off between compression and preserving the information that is relevant for the task Y.

3. On the Importance of Minimal Sufficient Representations

To formulate our hypothesis, we assume that the data of a modality are composed of an essence, which is all that information that can be found in both modalities; and nuisances, which refers to all that information that can be found only in one modality. In addition, our goal is to obtain representations of the inputs of each modality. Next, we explain more in detail these concepts. All the proofs of the Lemmas and Theorems can be found in Appendix A

3.1. Input Data: Essence and Nuisances

Definition 1. Let (X_α, X_β) be a pair of positive inputs from modalities α and β, respectively. We call *essence* to a variable Y that satisfies the following Markov Chains:

$$X_\beta \leftrightarrow Y \leftrightarrow X_\alpha \tag{4}$$
$$Y \leftrightarrow X_\beta \leftrightarrow X_\alpha \tag{5}$$

i.e., it refers to the common part of a positive pair of data. Although there exists more than one essence (infinite, in fact), all of them are equivalent under a one-to-one transformation, which is formalized next.

Lemma 1. *Let Y and Y' be essences of the same pair of modalities. Then, there exist a one-to-one transformation Ψ such that $Y = \Psi(Y')$.*

Equivalently, the partitions of X_α and X_β created by Y are unique. We note that the essence Y is a variable that we define to help us with the formulation, but our goal is not to discover Y itself, but the partition of the input set that it creates. For example, if we have two images with the same caption, we will consider that both images are equivalent in the sense that they belong to a common subset of the images set, but we are not interested in defining the subset. Thus, from now on we will refer to the essence as if it were a unique variable.

Definition 2. Let X_α be an input of modality α, Y the essence of X_α with respect to another modality. We call *nuisance* of modality α to a variable N_α that satisfies:

$$I(Y; N_\alpha) = 0 \tag{6}$$
$$I(X_\alpha; N_\alpha) = H(X_\alpha) - H(Y) = H(N_\alpha) \tag{7}$$

i.e., it refers to all information from X_α that cannot be found in the other modality.

Diagram in Figure 2 schematizes the relationships between all the elements described in this section.

3.2. Representations

Definition 3. We say that a variable Z_α is a *representation* of an input X_α if Z_α is a stochastic function of X_α or, equivalently, if Z_α is fully defined by $p(z_\alpha|x_\alpha)$.

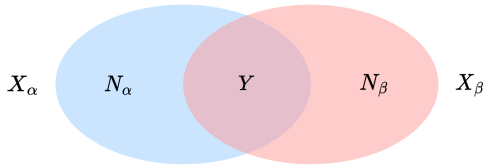


Figure 2: Diagram of the inputs, essence and nuisances

Given a representation Z_α of X_α, the following Markov chains are satisfied:

$$Y \leftrightarrow X_\alpha \leftrightarrow Z_\alpha \tag{8}$$
$$N_\alpha \leftrightarrow X_\alpha \leftrightarrow Z_\alpha \tag{9}$$

The goal of representation learning is to obtain representations with desirable properties for the problem to be solved. In the case of multimodal learning we consider two properties to be desirable: (i) sufficiency and (ii) minimality (Achille & Soatto, 2018a;b). We define these properties and argue their desirability next.

Definition 4. Given a representation Z_α of the input X_α and an essence Y. We call Z_α *sufficient* if it satisfies:

$$I(Z_\alpha; Y) = I(X_\alpha; Y) \tag{10}$$

i.e., a representation is sufficient if it preserves the essence in its entirety or, equivalently, if it preserves all the information that is common to both modalities. Because of Equation (8) we know by the Data Processing Inequality (DPI) that $I(Z_\alpha; Y) \leq I(X_\alpha; Y)$, i.e. $I(X_\alpha; Y)$ is an upper bound of $I(Z_\alpha; Y)$. Thus, the objective to be optimized to obtain a sufficient representation is:

$$\max_{Z_\alpha} I(Z_\alpha; Y) \tag{11}$$

Informally, *sufficiency is connected to the performance of our representations in downstream tasks*. Our representations must have all the information in the essence to perfectly solve *all* the tasks that can be derived from the essence. Here, we assume that tasks that are not in the essence cannot be solved. For example, if we have a set of images of dogs and a set of captions describing different aspects of them except the color, we cannot expect the text encoder to be able to *understand* the word "brown" even if there are brown dogs in our set of images. We formalize this next.

Theorem 1. *Let Y and Z_α be the essence and a representation of input X_α respectively, and let $\mathcal{T} = \{T : T = f(Y)\}$ be the set of all the deterministic functions of Y (i.e., all the tasks derived from Y). Then, we have that:*

$$p(t|z_\alpha) = p(t|x_\alpha)\ \forall\, T \in \mathcal{T} \implies I(Z_\alpha; Y) = I(X_\alpha; Y)$$

Definition 5. Given a representation Z_α and the nuisances N_α of an input X_α. We call Z_α *minimal* if it satisfies:

$$I(Z_\alpha; N_\alpha) = 0 \tag{12}$$

i.e., a representation is minimal if it eliminates all the nuisances of its modality or, in other words, if all the information that it contains can also be found in the input of the other modality. Since mutual information is non-negative, the objective to be optimized in order to obtain a minimal representation is:

$$\min_{Z_\alpha} I(Z_\alpha; N_\alpha) \tag{13}$$

Informally, *minimality is connected to representational alignment*. As explained in section 2, we can have representations with a good performance in a wide variety of downstream tasks but misaligned. Intuitively, even if two representations Z_α and Z_β have all the information about the essence (i.e., they are sufficient), if they also contain information about nuisances (i.e., they are not minimal), then the information they are encoding is different and, consequently, they will be misaligned. Revisiting the previous example, the sufficient representations of an image of *a **yellow** Labrador Retriever* and of an image of *a **brown** Labrador Retriever* could be different from each other, since they can be coding different information. However, the sufficient representations corresponding to the captions of each image will be presumably the same, since both images have presumably the same caption. Therefore, the presence of nuisances in the representations could cause misalignment, as stated next.

Theorem 2 (Informal)**.** *Let Z_α and Z_β be the representation of some inputs with nuisances N_α and N_β, respectively, such that Z_α and Z_β are aligned in the sense of equation (2). Then, $I(Z_\alpha; N_\alpha) = I(Z_\beta; N_\beta) = 0$.*

Summarizing the above, we have that our ideal representation should be a minimal sufficient statistic for Y. Equivalently, it should contain *only* (minimal) *all* (sufficient) the information that is common to both modalities (a.k.a. the essence). In this scenario, similarly to Lemma 1, we know that all the ideal representations create the same partition of the input. This connects to the "Anna Karenina principle" [1] that has been mentioned in different works in representation learning to hypothesize that all the well-performing models learn roughly the same internal representations (Bansal et al., 2021; Huh et al., 2024). Here, all the minimal sufficient ("*happy*") representations ("*families*") create the same partition of the input ("*are alike*").

4. Obtaining Minimal Sufficient Representations

We have discussed in the previous section the importance of minimal sufficient representations for good performance and alignment. We describe next a method to obtain them, which connects to the Information Bottleneck principle.

4.1. Obtaining Sufficient Representations

Equation (11) shows that $I(Z_\alpha; Y)$ must be maximized to find a representation Z_α that is sufficient. Since the essence Y is a variable that we have defined for our formulation whose distribution is unknown, calculating this term could seem problematic. However, in Appendix A.4 we show that $I(Z_\alpha; Y) = I(Z_\alpha; X_\beta)$. Thus, our objective becomes

$$\max_{Z_\alpha} I(Z_\alpha; X_\beta) \tag{14}$$

Computing exactly $I(Z_\alpha; X_\beta)$ is in general intractable, since it involves integrating over the entire space of β-modality inputs. However, we can obtain a lower bound of $I(Z_\alpha; X_\beta)$: since Z_β is a representation of X_β, we have that $Z_\alpha \leftrightarrow X_\beta \leftrightarrow Z_\beta$ and, by the DPI, it follows that $I(Z_\alpha; Z_\beta) \leq I(Z_\alpha; X_\beta)$. That is, $I(Z_\alpha; Z_\beta)$ is a lower bound of $I(Z_\alpha; Y)$. Analogously, $I(Z_\beta; Z_\alpha) \leq I(Z_\beta; Y)$, so given the symmetry of the mutual information, we must maximize $I(Z_\alpha; Z_\beta)$ in order to jointly maximize $I(Z_\alpha; Y)$ and $I(Z_\beta; Y)$. Thus, the objective becomes:

$$\max_{Z_\alpha, Z_\beta} I(Z_\alpha; Z_\beta) \tag{15}$$

Again, computing exactly $I(Z_\alpha; Z_\beta)$ requires integrating over the representation spaces, which is in general intractable. However, as explained in section 2, minimizing InfoNCE loss is equivalent to maximizing a lower bound of $I(Z_\alpha; Z_\beta)$. Thus, *encoders optimized through* InfoNCE *tend to give sufficient representations*. However, the resulting representations are not necessarily minimal due to the fact that this loss imposes no conditions on $I(Z_\alpha; N_\alpha)$. We derive in the next section a term that aims to increase the degree of minimality of the representations.

4.2. Obtaining Minimal Representations

Equation (13) shows that $I(Z_\alpha; N_\alpha)$ must be minimized to obtain a representation Z_α that is minimal. Since N_α is an abstract concept whose distribution is unknown, calculating this term could seem problematic. However, by the DPI and equation (9), we have that $I(Z_\alpha; N_\alpha) \leq I(Z_\alpha; X_\alpha)$. Thus, the objective to obtain a minimal representation becomes:

$$\min_{Z_\alpha} I(Z_\alpha; X_\alpha) \tag{16}$$

Again, computing exactly $I(Z_\alpha; X_\alpha)$ requires integrating over the representation and input spaces, which is intractable. However, we demonstrate in Appendix A.5 that:

$$I(Z_\alpha; X_\alpha) \leq \mathop{\mathbb{E}}_{p(x_\alpha, x_\beta)} \left[D_{KL} \left(p_{\theta_\alpha}(z|x_\alpha) || p_{\theta_\beta}(z|x_\beta) \right) \right] \tag{17}$$

[1] "All happy families are alike; each unhappy family is unhappy in its own way." (Tolstoy, 1877). This principle was popularized in (Diamond & Ordunio, 1999) to illustrate why only a small number of wild animals have been successfully domesticated over the course of history.

Therefore, we can minimize the given upper bound to minimize $I(Z_\alpha; X_\alpha)$. That is, the distributions of the representations of a positive pair of data from different modalities should be as similar as possible. Intuitively, if Z_α and Z_β are equal, then they can be affected only by the essence but not by the nuisances.

Spherical Gaussian Case The upper bound of equation (17) does not have a closed form in general. However, it is common to assume in practice that the representations distributions given the input are Gaussian, in which case, KL Divergence becomes tractable. In the case in which $p_{\theta_\alpha}(z|x_\alpha) = \mathcal{N}\left(z; \mu_{\theta_\alpha}(x_\alpha), \sigma^2 I\right)$ and $p_{\theta_\beta}(z|x_\beta) = \mathcal{N}\left(z; \mu_{\theta_\beta}(x_\beta), \sigma^2 I\right)$, as shown in Appendix A.6, we reach:

$$\begin{aligned} &\mathop{\mathbb{E}}_{p(x_\alpha, x_\beta)} \left[D_{KL}\left(p_{\theta_\alpha}(z|x_\alpha) || p_{\theta_\beta}(z|x_\beta) \right) \right] \propto \\ &\mathop{\mathbb{E}}_{p(x_\alpha, x_\beta)} \left[\left\| \mu_{\theta_\alpha}(x_\alpha) - \mu_{\theta_\beta}(x_\beta) \right\|_2^2 \right] = \mathcal{L}_M \end{aligned} \tag{18}$$

4.3. Information Bottleneck for two Modalities

Combining equations (14) and (16), the objective to obtain a representation Z_α that is sufficient and minimal becomes:

$$\max_{Z_\alpha} I(Z_\alpha; X_\beta) - \beta I(Z_\alpha; X_\alpha) \tag{19}$$

This is equivalent to an information bottleneck in which the task is the input of the other modality X_β. That is, Z_α must retain *only all* the information that is common between X_α and X_β. The same applies for Z_β. Combining equations (15) and (18), we have that this is equivalent to minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \beta \mathcal{L}_M \tag{20}$$

5. Toy Experiment

The objectives of this experiment are to empirically validate the different statements made throughout the previous sections and understand the relations between the different elements of our formulation. For this purpose, we use some datasets typically employed in disentanglement related tasks (Wang et al., 2024). Concretely, DSprites (Matthey et al., 2017), MPI3D (Gondal et al., 2019) and Shapes3D (Burgess & Kim, 2018) are used. These datasets contain images and labels that represent multiple independent factors of variation. We jointly train an image encoder and a factors encoder (i.e., images and factors are the two modalities). The reason to use these datasets is that we can control the amount of factors that we input to the encoder, thus controlling the information imbalance between both modalities. Unless otherwise stated, a ResNet20 (He et al., 2016) is used as image encoder, an MLP as encoder for the factors [2] and temperature in the InfoNCE loss is a trainable parameter initialized to 0.07. More details are given in Appendix C.

[2]We encode the factors using one-hot.

5.1. Does the contrastive loss alone remove nuisances?

To answer this question we propose several scenarios per dataset. In each scenario we provide all but one factor to the encoder, i.e., the nuisances of the image modality N_α are that missing factor. Thus, if the contrastive loss were eliminating the nuisances, then the image representations Z_α should contain no information about N_α, i.e., $I(Z_\alpha; N_\alpha) = 0$. We calculate a lower bound of this mutual information $\hat{I}(Z_\alpha, N_\alpha)$ by training a linear classifier from Z_α to N_α, following (Xu et al., 2020). We show in Table 1 the values of $\frac{\hat{I}(Z_\alpha; N_\alpha)}{H(N_\alpha)}$ for each dataset and category of factors[3]. This value encodes a lower bound of the ratio of the uncertainty of N_α that is reduced by observing Z_α and its value ranges from 0 to 1 (we call it uncertainty reduction ratio or simply URR), so if the contrastive loss alone removed all the nuisances, its value would be zero. We can extract the following conclusions from this: (i) a non-negligible amount of information about the missing factors is present in the image representation for every category; (ii) image encoder preserves more information on some categories than on others; and (iii) some categories are almost equally conserved among the datasets. We hypothesize that the last two points could be due to the inductive biases of the convolutional architecture of the image encoder (Cohen & Shashua, 2016; Mitchell, 2017; Wang & Wu, 2023), but exploring this point is out of the scope of this work.

Table 1: URR (in percentages) for each dataset and category. Some factors are not used because they do not fall into any category.

	DSprites	MPI3D	Shapes3D
Location	16.1 ± 3.7	12.4 ± 7.3	8.5 ± 0.1
Shape	77.1 ± 4.6	10.3 ± 1.0	8.7 ± 0.4
Size	66.3 ± 3.1	37.8 ± 0.9	7.2 ± 0.5
Objects Color	–	68.8 ± 2.3	54.1 ± 2.0

Not all architectures remove nuisances to the same extent It is well established that different neural architectures introduce distinct inductive biases (Raghu et al., 2021). Consequently, the extent to which nuisance factors are retained in the learned representations can vary depending on the model architecture. To investigate this, we replicate the previous experiment using a small Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the image encoder. Table 2 reveals two key observations: (i) local information—particularly *Location*—is less preserved in ViTs, likely due to their global attention mechanisms favoring long-range dependencies; and (ii) more global features—such as *Color*—are comparably preserved in both convolutional and transformer-based models. We emphasize that these trends may also depend

[3]We organize the factors into categories for ease of understanding of the conclusions. Information on what factors each category is composed of is provided in Appendix C.

on other architectural choices, such as the encoder depth, as explored in the following paragraph.

Table 2: URR (in percentages) for each dataset and category for ViT-based encoder.

	DSprites	MPI3D	Shapes3D
Location	2.8 ± 0.6	2.8 ± 0.3	1.1 ± 0.1
Shape	64.9 ± 1.4	7.0 ± 0.4	8.7 ± 0.2
Size	30.7 ± 3.5	20.8 ± 3.5	6.9 ± 1.5
Objects Color	–	63.5 ± 9.8	53.5 ± 1.6

Deeper neural encoders remove more nuisances It has been argued that the success of Deep Learning can be explained through the fact that deep neural networks implicitly introduce an Information Bottleneck (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). Intuitively, deterministic layers tend to remove information from the input because of the DPI. Thus, when the number of layers grows, the output of the neural network tends to be a more pruned version of the input but that preserves the information that is necessary to solve different downstream tasks (Alemi et al., 2016). We hypothesize then that the use of deeper neural encoders will tend to remove more nuisances. Effectively, we can observe in Figure 3 a trend among different factors in which deeper encoders remove more modality specific information. This can serve as an explanation for *The Capacity Hypothesis* stated in (Huh et al., 2024), which says that bigger models are more likely to converge to a shared representation than smaller models. We hypothesize that this representation is shared because it contains little information about the nuisances.

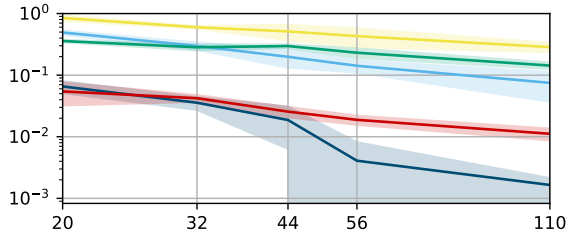


Figure 3: URR (y-axis) for different number of layers of the image neural encoder (x-axis). Same legend as Figure 4.

Higher temperatures remove more nuisances Wang & Liu (2021) observed that the value of the temperature in the InfoNCE loss considerably impacts on the entropy level of the representations. As stated in section 3.2, alignment is closely related to the level of nuisances in the representations and, consequently, to their entropy. We run an experiment identical to the previous one for some factors in which the temperature is fixed. Its results are shown in Figure 4 and we observe that (i) higher values of temperature tend to remove more nuisances and (ii) not all the factors are equally affected by the changes in temperature.

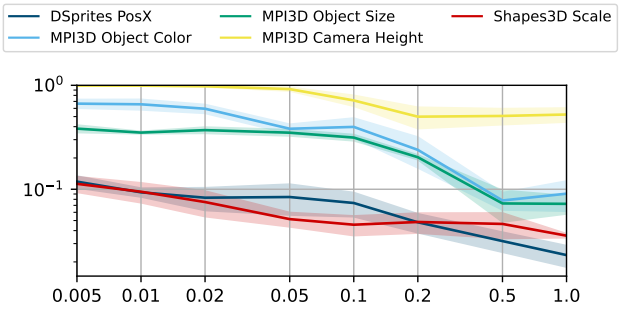


Figure 4: URR (y-axis) for different values of temperature (x-axis).

5.2. Does the presence of nuisances in the representation negatively correlate with the level of alignment?

As informally demonstrated in section 3.2, the fact that two representations are minimal is a necessary condition for them to be aligned. We hypothesize that misalignment is just an effect of an information imbalance in the representation space. To empirically analyze this phenomenon, we design an experiment similar to the previous one. In this case, more than one factor can be removed, i.e., N_α can be a set of factors. We generate 100 scenarios per dataset in which a randomly chosen subset of factors N_α is not provided as input to the factors encoder. Similarly to the previous experiment, we calculate $\hat{I}(Z_\alpha; N_\alpha)$ and the CKA metric. In Figure 5 it is shown that, for the three datasets, there exists a negative correlation between $\hat{I}(Z_\alpha; N_\alpha)$ and the alignment value.

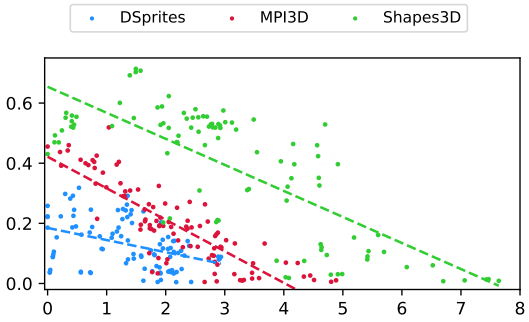
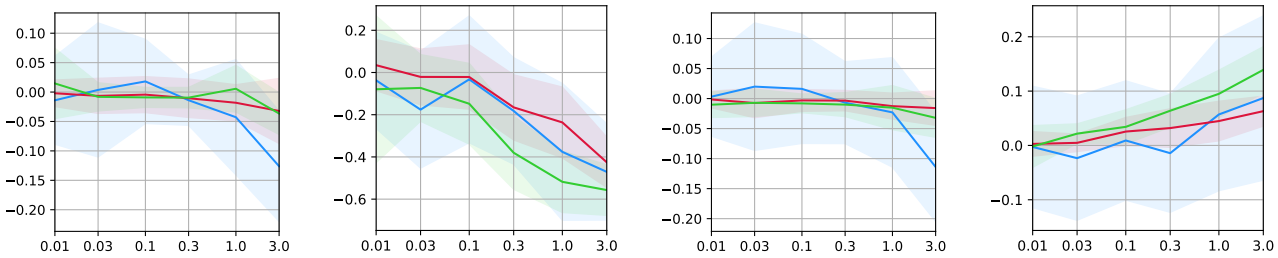


Figure 5: Alignment (y-axis) vs. $\hat{I}(Z_\alpha; N_\alpha)$ (x-axis).

5.3. Does our regularization term effectively increase the alignment level?

To analyze this, we randomly select, for each dataset, 10 of the 100 scenarios above and we train the encoders for different values of β. In all of them we set the temperature fixed to 0.01. We show in Figure 6 the value of different measures for different values of β. We can extract the following conclusions from this: (i) lower values of β retain better the essence (Fig. 6a), since increasing $I(Z_\alpha; Y)$ prevails over decreasing $I(Z_\alpha; N_\alpha)$; (ii) lower values of β also tend to retain more nuisances, since more entropic representations are encouraged in this case (Fig. 6b); (iii) higher values of

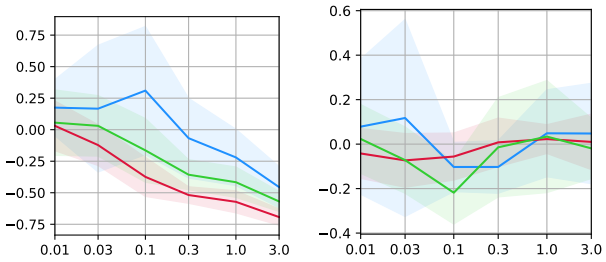


(a) $\hat{I}(Z_\alpha; Y)$ (b) $\hat{I}(Z_\alpha; N_\alpha)$ (c) Accuracy of Y (d) Alignment

Figure 6: Relative change (with respect to the case in which $\beta = 0$) of different measures (y-axis) vs. β (x-axis). The temperature is equal to 0.01 in all the cases. Same legend as Figure 5.

β remove more nuisances, since they favor the decreasing of $I(Z_\alpha; N_\alpha)$ (Fig. 6b); (iv) higher values of β also tend to discard more information of the essence, since they promote less entropic representations (Fig. 6a); (v) representations with lower $I(Z_\alpha; Y)$ result in a lower accuracy (Figs. 6a and 6c), as stated in Theorem 1; and (vi) representations with lower $I(Z_\alpha; N_\alpha)$ result in a higher alignment (Figs. 6b and 6d), as stated in Theorem 2.

On the *Information Homeostasis* of the representations In the previous experiment the temperature has been set fixed. However, as shown in Figure 4, lower temperatures tend to preserve more information of the nuisances. Thus, the next question arises: *Will the temperature be affected by the value of β?* To answer it, we repeat the previous experiment but setting the temperature as a trainable parameter. We show the results in Figure 7 and we observe that: (i) temperature tends to become lower when higher values of β are used (Fig. 7a); and (ii) this translates into the fact that nuisances tend not to be eliminated to the same extent as in the case in which the temperature is fixed (Fig. 6b vs. 7b). We call this phenomenon *Information Homeostasis*, since it seems that, when an external stimulus (i.e., increasing β) affects the encoder, this activates available mechanisms (i.e., decreasing the temperature) in order to preserve to the extent possible the entropy of the representations (DelMonte & Kim, 2011). This effect becomes more pronounced for the highest values of β. In these cases, the level of nuisances is similar to the case in which $\beta = 0$, which reminds of an *homeostatic range* (Kotas & Medzhitov, 2015). This is an intriguing phenomenon that is out of the scope of this work.



(a) Temperature (b) $\hat{I}(Z_\alpha; N_\alpha)$

Figure 7: Relative change (with respect to the case in which $\beta = 0$) of different measures (y-axis) vs. β (x-axis). Temperature is a trainable parameter. Same legend as Figure 5.

6. Real-World Applications

Several real-world applications benefits from aligned representations. These include those that consist in generating one modality from another. We analyze what are the implications of introducing our regularization term in a real-world scenario. Concretely, we train a Q-Former with a frozen decoder-based LLM (Li et al., 2023) with different loss functions. In all the cases, two terms are present: (i) an image-text contrastive loss (ITC), i.e., the InfoNCE loss between image and text representations; and (ii) a language model loss (LM), which trains the Q-Former to generate text through the LLM using images as the condition. However, none of these losses explicitly encourages a high representational alignment. Thus, we experiment by adding: (i) an image-text matching loss (ITM), which is the binary cross-entropy loss of a task in which the model must predict if an image-text pair is positive or negative (Chen et al., 2020b); or (ii) our regularization term in equation (20). We note that, in contrast to ITM, our regularization term is modality-agnostic, computationally efficient and straightforward to implement. COCO (Lin et al., 2014) is used to train and test our model. More details are given in Appendix C.

6.1. Image Captioning

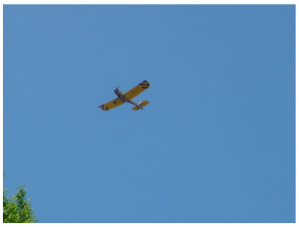
We argue that, for optimal performance in image captioning, the learned image representations should contain as little information as possible about nuisance factors. When nuisance information is retained, representations may encode fine-grained visual details that the text decoder is not equipped to handle, as it has not been trained to exploit such information. Table 3 summarizes the performance of different models. We observe the following trends: (i) loss functions that promote alignment between modalities tend to improve image captioning performance; (ii) there is a trade-off between image captioning and retrieval: caption-

Table 3: CIDEr (Vedantam et al., 2015), BLEU@4 (Papineni et al., 2002) and retrieval accuracy for Q-Formers trained with different loss functions.

	CIDEr	BLEU@4	I2T R@1	T2I R@1
ITC+LM	91.7 ± 0.2	28.6 ± 0.1	$\mathbf{64.2 \pm 0.2}$	$\mathbf{52.3 \pm 0.4}$
ITC+LM+ITM	91.8 ± 0.5	28.8 ± 0.2	61.4 ± 0.6	49.7 ± 0.8
ITC+LM+$0.01\mathcal{L}_M$	92.3 ± 0.8	29.1 ± 0.4	64.0 ± 0.3	$\mathbf{52.3 \pm 0.5}$
ITC+LM+$0.03\mathcal{L}_M$	92.6 ± 0.3	29.2 ± 0.2	63.9 ± 0.4	52.1 ± 0.5
ITC+LM+$0.1\mathcal{L}_M$	$\mathbf{93.0 \pm 0.3}$	$\mathbf{29.4 \pm 0.3}$	63.0 ± 0.5	50.4 ± 0.5
ITC+LM+$0.3\mathcal{L}_M$	90.5 ± 0.4	28.5 ± 0.2	59.6 ± 0.4	47.1 ± 0.4

ing benefits from minimal representations, whereas sufficient representations favors retrieval (as it is a downstream task); (iii) our loss slightly improves captioning performance for low values of β, with minimal degradation in retrieval performance; (iv) for moderate values of β, our loss substantially enhances captioning performance, though at a higher cost to retrieval accuracy; and (v) at high values of β, captioning and retrieval performances drop sharply, as representations become overly compressed and fail to retain sufficient task-relevant information—mirroring the trend found in Figure 6.

We also show in Figure 8 and in Appendix D some of the captions generated by the different models in Table 3. We observe that, in those cases in which representations are less aligned, captions tend to be more entropic because the representations are as well. This, in some cases, translates into captions that have information that does not correspond to the image. From a geometric point of view, we believe that in a misaligned space, for example, the representations of "lots of trees" and "a tree" are closer in the image than in the text space and, thus, the text decoder "confuses" them.



(1): a small airplane flying over a forest with lots of trees
(2): a small airplane flying through a blue sky above trees
(3): a small plane flying through a blue sky above a tree
(5): a small plane flying through a blue sky above a tree

(1): a cat is laying down on a bed with a white blanket
(2): a black and white cat laying on a red blanket
(3): a black and white cat laying on a red blanket
(5): a black wnd white cat laying on a bed

Figure 8: Captions generated by some of the trained models. Numbers correspondence is the same as in Table 3.

6.2. Multimodal Representation Space Arithmetic

Figure 9 shows image retrievals obtained from combining image and text representations from the Q-Former trained with $\beta = 0.01$. Examples including other loss functions are found in Appendix E, showing that those not encouraging alignment, result in worse multimodal retrievals.

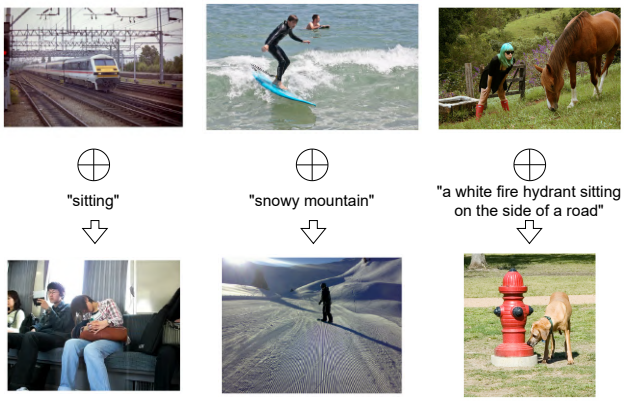


Figure 9: Multimodal image retrieval for $\beta = 0.01$.

7. Related Work

Information Bottleneck and Contrastive Representation Learning Other works have previously connected IB to CRL, especially in the context of multi-view learning. In (Tian et al., 2020b), the authors argue that in CRL good views for a given task are those that optimize an IB w.r.t. that given task. Federici et al. (2020), analogously to us, propose a loss function to obtain representations that retain only the information shared by the two views, which is considered to be the relevant for downstream tasks. Tsai et al. (2020) build on the previous one and argue that including a reconstruction loss encourages to preserve the downstream task-relevant information. In (Wang et al., 2022) it is argued that, in the multi-view setting, imposing a strong IB can be detrimental for downstream tasks, since it could be removing an excess of information in the representation. However, to the best of our knowledge, the use of the IB in the multimodal setting remains understudied and the present is the first work that explores this and connects it to the misalignment typically present between representations from different modalities.

Alignment of Multimodal Contrastive Learned Representations This is a well-studied field but still marked by a great amount of unanswered questions. The first work that extensively studied the alignment in CRL was (Wang & Isola, 2020) and demonstrated that, under infinite negative samples, the InfoNCE is globally minimized if the representations are perfectly aligned. They also define the representational alignment as in equation (17), which makes our formulation consistent with this work. However, it was observed in (Liang et al., 2022) that there exists in practice a great misalignment between representations from different modalities. This misalignment becomes problematic for tasks that need to combine both modalities. For example, Chen et al. (2020b); Li et al. (2021; 2022; 2023) use modifications to the InfoNCE loss to obtain a better performance in tasks in which image representations serve to obtain text, such as Image Captioning or Visual Question Answering. In the the opposite direction, Ramesh et al. (2022) use a generative model to transform text representations to image representations to train a text-conditioned image generator. In our view, this generative model serves to increase the entropy of the text representations to balance them with the image representations, which are typically more entropic. To the best of our knowledge, (Schrodi et al., 2024) is the first work in which this phenomenon is explained through the lens of an information imbalance. However, this imbalance is analyzed in the input space rather than in the representation space. Thus, aspects such as the encoder depth and hyperparameters or modifications of the loss function are not explored.

8. Conclusions

We give an explanation to the phenomenon of multimodal misalignment that usually emerges in encoders trained to minimize contrastive losses. These are designed to obtain representations that preserve the information about what is common to both modalities, but not to remove modality-specific information. We theoretically and empirically show that the presence of this modality-specific information in the representations is correlated with the misalignment phenomenon. We also examine the impact that different hyperparameters such as the temperature or encoder depth have on how much of this modality-information is removed. We derive a term that can be added to the contrastive loss which aims to eliminate this modality-specific information and, thus, allows to obtain a more aligned representation space. We find a phenomenon that we call *Information Homeostasis*, which consists in the fact that encoders seem to prefer representations with more nuisances and they modify, if possible, some of their internal parameters for this purpose. Finally, we show that our term in the loss function translates into a better performance in image captioning and seems to result in more consistent multimodal image retrievals.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007666, MCIN/AEI/10.13039/501100011033 under Grant PID2021-126061OB-C44, and the Government of Aragón (Grant Group T36 23R). We are also grateful to the French National Research Agency for their support through the ANR-20-CE23-0012-01 (MIM) grant.

This work originated during the JSALT 2024 workshop. We gratefully acknowledge the researchers, mentors, and collaborators who took part in the workshop, as well as the staff at the Center for Language and Speech Processing at Johns Hopkins University, for cultivating a collaborative and intellectually rich environment that was instrumental in shaping this research. The workshop was partially supported by generous contributions from Amazon, Facebook, Google, and Microsoft.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018a.

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018b.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Burgess, C. and Kim, H. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020b.

Cohen, N. and Shashua, A. Inductive bias of deep convolutional networks through pooling geometry. *arXiv preprint arXiv:1605.06743*, 2016.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Cuervo, S. and Marxer, R. Scaling properties of speech language models. *arXiv preprint arXiv:2404.00685*, 2024.

DelMonte, D. W. and Kim, T. Anatomy and physiology of the cornea. *Journal of Cataract & Refractive Surgery*, 37 (3):588–598, 2011.

Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Diamond, J. M. and Ordunio, D. *Guns, germs, and steel*, volume 521. Books on Tape New York, 1999.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fahim, A., Murphy, A., and Fyshe, A. Its not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Gondal, M. W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*, 2023.

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., and Luo, Y. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Kotas, M. E. and Medzhitov, R. Homeostasis, inflammation, and disease susceptibility. *Cell*, 160(5):816–827, 2015.

Lee, G.-G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., Guo, S., Wu, Z., Liu, Z., Wang, H., et al. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*, 2023.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Mitchell, B. R. *The spatial inductive bias of deep learning*. PhD thesis, Johns Hopkins University, 2017.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Ramasinghe, S., Shevchenko, V., Avraham, G., and Thalaiyasingam, A. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27263–27272, 2024.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Schrodi, S., Hoffmann, D. T., Argus, M., Fischer, V., and Brox, T. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *arXiv preprint arXiv:2404.07983*, 2024.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Sinha, G., Shahi, R., and Shankar, M. Human computer interaction. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pp. 1–4. IEEE, 2010.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020a.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020b.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Udandarao, V. Understanding and fixing the modality gap in vision-language models. *Master's thesis, University of Cambridge*, 2022.

Usunier, N., Buffoni, D., and Gallinari, P. Ranking with ordered weighted pairwise classification. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1057–1064, 2009.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Wang, F. and Liu, H. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

Wang, H., Guo, X., Deng, Z.-H., and Lu, Y. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16041–16050, 2022.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wang, X., Chen, H., Wu, Z., Zhu, W., et al. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wang, Z. and Wu, L. Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems*, 36:74289–74338, 2023.

Weston, J., Bengio, S., and Usunier, N. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81:21–35, 2010.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., and López, A. M. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):537–547, 2020.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

Yager, R. R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190, 1988.

A. Proofs of Sections 3 and 4

A.1. Proof of Lemma 1

Lemma 1. *Let Y and Y' be essences of the same pair of modalities. Then, there exist a one-to-one transformation Ψ such that $Y = \Psi(Y')$.*

Proof. By Definition 1, Y and Y' are minimal sufficient statistics of X_α for X_β. Then, by the definition of minimal sufficient statistic, there exist two functions Ψ and Ψ' such that $Y = \Psi(Y')$ and $Y' = \Psi'(Y)$. Then, $\Psi^{-1} = \Psi'$. □

A.2. Proof of Theorem 1

Theorem 1. *Let Y and Z_α be the essence and a representation of input X_α respectively, and let $\mathcal{T} = \{T : T = f(Y)\}$ be the set of deterministic functions of Y (i.e., all the tasks derived from Y). Then, we have that:*

$$p(t|z_\alpha) = p(t|x_\alpha) \; \forall \, T \in \mathcal{T} \implies I(Z_\alpha; Y) = I(X_\alpha; Y)$$

Proof. First, we know from equation (4) that $I(Y; Z_\alpha) \leq I(Y; X_\alpha)$. Second, since $T = f(Y)$, we have the Markov Chain $Z_\alpha \leftrightarrow Y \leftrightarrow X_\alpha$ and, thus, by the DPI, $I(T; Z_\alpha) \leq I(Y; Z_\alpha)$. Third, since $p(t|x_\alpha) = p(t|z_\alpha)$, we know that $I(T; X_\alpha) = I(T; Z_\alpha)$. Thus, we have that $I(T; X_\alpha) = I(T; Z_\alpha) \leq I(Y; Z_\alpha) \leq I(Y; X_\alpha)$. Finally, since T can be any function of Y, it can be the identity function, in which case $I(T; X_\alpha) = I(Y; X_\alpha)$. □

A.3. Proof of Theorem 2

Theorem 2 (Informal)**.** *Let Z_α and Z_β be two representations of a pair of inputs with nuisances N_α and N_β respectively, such that Z_α and Z_β are aligned in the sense of equation (2). Then, $I(Z_\alpha; N_\alpha) = I(Z_\beta; N_\beta) = 0$.*

Proof. First, if $I(Z_\alpha; N_\alpha) \neq 0$, then there exists a surjective function f such that $Z_\beta = f(Z_\alpha)$, i.e., more than one representation of modality α can correspond with one representation of modality β. Let $\left\{ z_\alpha^{(l)} \sim p_{\theta_\alpha}\left(z|x_\alpha^{(l)}\right) : x_\alpha^{(l)} \sim p(x_\alpha) \right\}$ and $\left\{ z_\beta^{(l)} \sim p_{\theta_\beta}\left(z|x_\beta^{(l)}\right) : x_\beta^{(l)} \sim p(x_\beta) \right\}$ be two infinite sets. Then, there exists a pair (l, l') for which $z_\alpha^{(l)} \neq z_\alpha^{(l')}$ and $z_\beta^{(l)} = z_\beta^{(l')}$ and for which $K(z_\alpha^{(l)}, z_\alpha^{(l)}) \neq K(z_\alpha^{(l)}, z_\alpha^{(l')})$, but $K(z_\beta^{(l)}, z_\beta^{(l)}) = K(z_\beta^{(l)}, z_\beta^{(l')})$. □

We must note that, in practice, we usually work with finite sets, so we could obtain the maximum value of alignment while having the presence of nuisances in the representation.

A.4. Proof of $I(Z_\alpha; Y) = I(Z_\alpha; X_\beta)$

Proof.

$$p(z_\alpha|y, x_\beta) = \int p(z_\alpha|y, x_\beta, x_\alpha) p(x_\alpha|y, x_\beta) \, dx_\alpha \tag{21}$$

$$= \int p(z_\alpha|x_\alpha) p(x_\alpha|y) \, dx_\alpha = p(z_\alpha|y) \tag{22}$$

In line (21), we apply Definition 3 and equation (4). Then, we have the following Markov Chain $X_\beta \leftrightarrow Y \leftrightarrow Z_\alpha$ and, by the DPI, $I(Z_\alpha; Y) \geq I(Z_\alpha; X_\beta)$.

$$p(z_\alpha|y, x_\beta) = \int p(z_\alpha|y, x_\beta, x_\alpha) p(x_\alpha|y, x_\beta) \, dx_\alpha \tag{23}$$

$$= \int p(z_\alpha|x_\alpha) p(x_\alpha|x_\beta) \, dx_\alpha = p(z_\alpha|x_\beta) \tag{24}$$

In line (23), we apply Definition 3 and equation (5). Then, we have the following Markov Chain $Y \leftrightarrow X_\beta \leftrightarrow Z_\alpha$ and, by the DPI, $I(Z_\alpha; Y) \leq I(Z_\alpha; X_\beta)$. □

A.5. Proof of equation (17)

Proof.

$$I(Z_\alpha; X_\alpha) = \iint p_{\theta_\alpha}(z, x_\alpha) \log \frac{p_{\theta_\alpha}(z|x_\alpha)}{p_{\theta_\alpha}(z)} \, dz \, dx_\alpha \tag{25}$$

$$= \iiint p_{\theta_\alpha}(z|x_\alpha) p(x_\alpha, x_\beta) \log \frac{p_{\theta_\alpha}(z|x_\alpha)}{p_{\theta_\alpha}(z)} \, dz \, dx_\alpha \, dx_\beta \tag{26}$$

$$= \iiint p_{\theta_\alpha}(z|x_\alpha) p(x_\alpha, x_\beta) \log \frac{p_{\theta_\alpha}(z|x_\alpha)}{p_{\theta_\beta}(z|x_\beta)} \frac{p_{\theta_\beta}(z|x_\beta)}{p_{\theta_\alpha}(z)} \, dz \, dx_\alpha \, dx_\alpha \tag{27}$$

$$= \mathop{\mathbb{E}}_{p(x_\alpha, x_\beta)} \left[D_{KL} \left(p_{\theta_\alpha}(z|x_\alpha) || p_{\theta_\beta}(z|x_\beta) \right) \right] - \mathop{\mathbb{E}}_{p(x_\alpha, x_\beta|z)} \left[D_{KL} \left(p_{\theta_\alpha}(z) || p_{\theta_\beta}(z|x_\beta) \right) \right] \tag{28}$$

$$\leq \mathop{\mathbb{E}}_{p(x_\alpha, x_\beta)} \left[D_{KL} \left(p_{\theta_\alpha}(z|x_\alpha) || p_{\theta_\beta}(z|x_\beta) \right) \right] \tag{29}$$

□

A.6. Proof of equation (18)

Proof. Let $p_{\theta_\alpha}(z \mid x_\alpha) = \mathcal{N}(z; \mu_\alpha, \sigma^2 I)$ and $p_{\theta_\beta}(z \mid x_\beta) = \mathcal{N}(z; \mu_\beta, \sigma^2 I)$, with $\mu_\alpha = \mu_{\theta_\alpha}(x_\alpha)$ and $\mu_\beta = \mu_{\theta_\beta}(x_\beta)$. The KL divergence between these two Gaussians is given by:

$$D_{\mathrm{KL}}(p \,\|\, q) = \frac{1}{2} \left[\mathrm{tr}(\Sigma_q^{-1} \Sigma_p) + (\mu_q - \mu_p)^\top \Sigma_q^{-1} (\mu_q - \mu_p) - d + \log \frac{\det \Sigma_q}{\det \Sigma_p} \right] \tag{30}$$

where $p = \mathcal{N}(\mu_p, \Sigma_p)$ and $q = \mathcal{N}(\mu_q, \Sigma_q)$. For $\Sigma_p = \Sigma_q = \sigma^2 I$, Eq. (30) simplifies to:

$$D_{\mathrm{KL}} \left(p_{\theta_\alpha}(z \mid x_\alpha) \,\|\, p_{\theta_\beta}(z \mid x_\beta) \right) = \frac{1}{2\sigma^2} \left\| \mu_{\theta_\alpha}(x_\alpha) - \mu_{\theta_\beta}(x_\beta) \right\|_2^2 . \tag{31}$$

Taking expectation over the joint distribution $p(x_\alpha, x_\beta)$, we obtain:

$$\mathbb{E}_{p(x_\alpha, x_\beta)} \left[D_{\mathrm{KL}} \left(p_{\theta_\alpha}(z \mid x_\alpha) \,\|\, p_{\theta_\beta}(z \mid x_\beta) \right) \right] = \frac{1}{2\sigma^2} \mathbb{E}_{p(x_\alpha, x_\beta)} \left[\left\| \mu_{\theta_\alpha}(x_\alpha) - \mu_{\theta_\beta}(x_\beta) \right\|_2^2 \right] \tag{32}$$

Hence, the expected KL divergence is proportional to the expected squared ℓ_2 distance between the mean embeddings:

$$\mathbb{E}_{p(x_\alpha, x_\beta)} \left[D_{\mathrm{KL}} \left(p_{\theta_\alpha}(z \mid x_\alpha) \,\|\, p_{\theta_\beta}(z \mid x_\beta) \right) \right] \propto \mathbb{E}_{p(x_\alpha, x_\beta)} \left[\left\| \mu_{\theta_\alpha}(x_\alpha) - \mu_{\theta_\beta}(x_\beta) \right\|_2^2 \right] . \tag{33}$$

The constant of proportionality is $\frac{1}{2\sigma^2}$ and independent of the model parameters $\theta_\alpha, \theta_\beta$. □

B. Connection between our loss function and temperature

In the case where the embeddings are unit-norm, our proposed loss takes the form:

$$\mathcal{L}_i = \log \frac{\exp(s_{ii}/\tau)}{\sum_k \exp(s_{ik}/\tau)} + 2\beta(1 - s_{ii}), \tag{34}$$

where s_{ik} denotes the cosine similarity between the embeddings $z^{(i)}$ and $z^{(k)}$. The gradient of $\mathcal{L}_i$ with respect to each similarity term is given by:

$$\frac{\partial \mathcal{L}_i}{\partial s_{ii}} = -\frac{1}{\tau}\left(1 - \frac{\exp(s_{ii}/\tau)}{\sum_k \exp(s_{ik}/\tau)}\right) - 2\beta, \tag{35}$$

$$\frac{\partial \mathcal{L}_i}{\partial s_{ij}} = \frac{1}{\tau} \cdot \frac{\exp(s_{ij}/\tau)}{\sum_k \exp(s_{ik}/\tau)} \quad \text{for } j \neq i. \tag{36}$$

We further analyze a modified variant of the InfoNCE loss where the temperature differs between the numerator and denominator:

$$\mathcal{L}'_i = \log \frac{\exp(s_{ii}/\tau')}{\sum_k \exp(s_{ik}/\tau)}. \tag{37}$$

The gradients of this variant are:

$$\frac{\partial \mathcal{L}'_i}{\partial s_{ii}} = -\frac{1}{\tau'}\left(1 - \frac{\exp(s_{ii}/\tau')}{\sum_k \exp(s_{ik}/\tau)}\right), \tag{38}$$

$$\frac{\partial \mathcal{L}'_i}{\partial s_{ij}} = \frac{1}{\tau} \cdot \frac{\exp(s_{ij}/\tau)}{\sum_k \exp(s_{ik}/\tau)} \quad \text{for } j \neq i, \tag{39}$$

which matches Eq. (36), confirming that the two losses only differ in their treatment of s_{ii}.

By comparing Eq. (35) and Eq. (38), we find that our regularized loss is equivalent to optimizing a variant of InfoNCE with a temperature mismatch between numerator and denominator. Solving for β, we obtain:

$$\beta = \frac{1}{2}\left[\frac{\tau - \tau'}{\tau\tau'} + \frac{\exp(s_{ii}/\tau) - \exp(s_{ii}/\tau')}{\sum_k \exp(s_{ik}/\tau)}\right]. \tag{40}$$

This expression reveals that:

- The effective temperature gap depends on the similarity between the anchor and all other samples in the batch.
- When $s_{ii} \ll \sum_k \exp(s_{ik}/\tau)$, i.e., predictions are far from the target distribution, the second term in Eq. (40) is negligible, and
$$\beta \approx \frac{1}{2} \cdot \Delta\tau$$
with $\Delta\tau = \frac{\tau - \tau'}{\tau\tau'}$. Thus, larger values of β correspond to larger temperature mismatches.
- Conversely, when $s_{ii} \approx \sum_k \exp(s_{ik}/\tau)$, i.e., the model is confident in its match, we have:
$$\beta \approx \frac{1}{2}\left[\Delta\tau + 1 - \exp(\Delta\tau)\right],$$
indicating that the temperature gap required to match a fixed β is smaller in this regime.

Hence, our regularizer can be interpreted as an adaptive temperature adjustment that decreases the denominator temperature when the model is uncertain, and aligns it closer to the numerator temperature when predictions are confident.

C. Experimental Details

Below, we provide a summary of the experimental setup. Full implementation details and code are available at: https://github.com/antonioalmudevar/multimodal_ib.

Table 4: Hyperparameters of Section 5

	DSprites	MPI3D	Shapes3D
factors encoder MLP	$\{16, 16, 8, 8, 16, 16\}$	$\{128, 128, 64, 64, 128, 128\}$	$\{64, 64, 32, 32, 64, 64\}$
number of epochs	50	50	50
batch size	128	128	128
optimizer	Adam	Adam	Adam
learning rate	0.001	0.001	0.001
scheduler	Step	Step	Step
step size (epochs)	20	20	20
scheduler γ	0.3	0.3	0.3

Table 5: Categories of factors in section 5.1. In MPI3D, the factor background_color actually refers to the color of a ring in the images, so we consider it as an object (see https://github.com/rr-learning/disentanglement_dataset). There are some factors missing because the do not fall into any category.

	DSprites	MPI3D	Shapes3D
Location	posX, posY	horizontal_axis, vertical_axis	orientation
Shape	shape	object_shape	shape
Size	size	object_size	scale
Object Color	-	object_color, background_color	object_hue

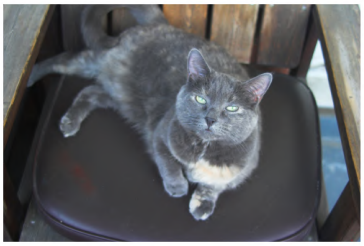
Table 6: Hyperparameters of Section 6

vision encoder	VIT-g/14 (Fang et al., 2023)
image size	224
# of query tokens	32
cross attention frequency	2
representation dimension	256
text encoder	BERT$_{base}$(Devlin, 2018)
batch size	128
optimizer	Adam
learning rate	0.0001
optimizer β	$(0.9, 0.999)$
scheduler	cosine annealing
warm-up steps	1000
training steps	50000

D. More Results of Section 6.1

(1): a woman sitting on a bench with a bag of food
(2): a woman sitting on a bench with a plate of food
(3): a woman sitting on a bench in the snow
(5): a woman sitting on a bench in the snow

(1): a man is taking a picture of himself in the mirror
(2): a man is taking a picture of himself in the mirror
(3): a man is seen in the reflection of a mirror
(5): a man standing in front of a bathroom mirror



(1): a cat laying on top of a wooden chair in a room
(2): a cat sitting on a chair looking at the camera
(3): a cat sitting on a chair looking at the camera
(5): a cat laying on top of a wooden chair

(1): a couple of benches that are in the grass
(2): a white park bench sitting next to a tree
(3): a bench sitting in the middle of a park
(5): a metal bench sitting in the middle of a park

(1): a toy set of a man in a hospital bed with a robot in the middle
(2): two dummy heads are on a bed that is shaped like a boat
(3): a fake bed with a dummy head and legs on it
(5): a display of a demonic joker character in a bed

(1): a man and woman in a white dress are sitting on a bed
(2): a man and woman are sitting in a bed
(3): a picture of a couple in a bedroom with a glass frame
(5): a photo of a couple in a frame on a bed

(1): a person holding a teddy bear with a pair of scissors
(2): a person holding a teddy bear with a woman's pregnant belly
(3): a person with a teddy bear on their lap
(5): a pregnant woman holding a teddy bear with a baby on it

(1): a jetliner flying through a cloudy blue sky
(2): a plane flying in the sky with a half moon in the background
(3): a plane flying in the sky with a half moon in the background
(5): a plane flying in the sky with a half moon in the background

(1): a group of horses standing around a baby horse
(2): a group of horses standing around a pile of hay
(3): a group of horses eating hay in a field
(5): a group of horses standing next to each other

(1): a couple of chairs and an umbrella on a field
(2): a couple of chairs sitting next to a table
(3): a couple of chairs and a table on a field
(5): a couple of chairs sitting next to a fence

(1): a man standing under a banner with a frisbee
(2): a man standing under a banner that says bicycle fair
(3): a man standing under a banner that says bicycle relief
(5): a man standing under a sign that reads for a cause

(1): a cow standing in a field with a house in the background
(2): a cow is standing in a field with a house in the background
(3): a couple of cows standing on top of a lush green field
(5): a cow is standing in a field with a few trees

(1): a dog herding sheep in a field with a dog
(2): a dog herding sheep in a field with a dog running behind them
(3): a dog is herding three sheep in a field
(5): a dog is herding some sheep in a field

(1): a wooden bench sitting in the middle of a yard
(2): a pile of wood sitting on the side of a road
(3): a pile of wood sitting on the side of a road
(5): a bunch of old used wooden poles in a yard

(1): a close up of a pair of scissors in a pile
(2): a group of scissors that are hanging up together
(3): a bunch of pink scissors are chained together in a room
(5): a bunch of pink scissors are all grouped up

Figure 10: Captions generated by some of the trained models. Numbers correspondence is the same as in Table 3.

E. More Results of Section 6.2

"a woman with her face close to a mans face"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$

ITC+LM+0.1$\mathcal{L}_M$

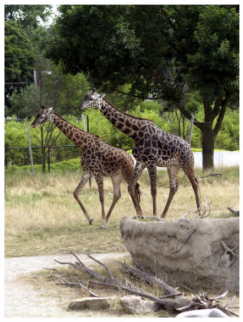
"a white fire hydrant sitting on the side of a road"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$

ITC+LM+0.1$\mathcal{L}_M$



⊕ "snowy mountain"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$

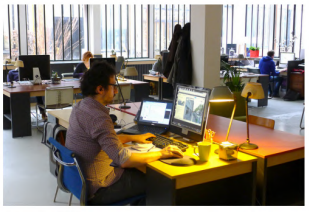
ITC+LM+0.1$\mathcal{L}_M$

⊕ "modern building"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$



ITC+LM+0.1$\mathcal{L}_M$

⊕ "old"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$

ITC+LM+0.1$\mathcal{L}_M$

⊕ "bike"

ITC+LM

ITC+LM+ITM

ITC+LM+0.01$\mathcal{L}_M$

ITC+LM+0.1$\mathcal{L}_M$

Figure 11: Multimodal image retrievals from models train with different loss functions. We believe that text representations from more simple captions (less entropic) are better aligned with image representations from encoders trained with a higher values of β, since they are less entropic too.