RELATING BY CONTRASTING: A DATA-EFFICIENT FRAMEWORK FOR MULTIMODAL DGMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning for generative models often refers to the learning of abstract concepts from the *commonality* of information in multiple modalities, such as vision and language. While it has proven effective for learning generalisable representations, the training of such models often requires a large amount of "related" multimodal data that shares commonality, which can be expensive to come by. To mitigate this, we develop a novel contrastive framework for generative model learning, allowing us to train the model not just by the commonality between modalities, but by the *distinction* between "related" and "unrelated" multimodal data. We show in experiments that our method enables data-efficient multimodal learning on challenging datasets for various multimodal variational autoencoder (VAE) models. We also show that under our proposed framework, the generative model can accurately identify related samples from unrelated ones, making it possible to make use of the plentiful unlabeled, unpaired multimodal data.

1 INTRODUCTION

To comprehensively describe concepts in the real world, humans collect multiple perceptual signals of the same object such as image, sound, text and video. We refer to each of these media as a *modality*, and a collection of different media featuring the same underlying concept is characterised as *multimodal* data. Learning from multiple modalities has been shown to yield more generalisable representations (Zhang et al., 2020; Guo et al., 2019; Yildirim, 2014), as different modalities are often complimentary in content while overlapping for common abstract concept.

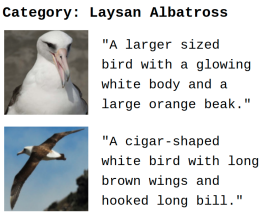
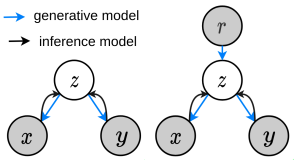


Figure 1: Multimodal data from the CUB dataset

Despite the motivation, it is worth noting that the multimodal framework is not exactly data-efficient—constructing a suitable dataset requires a lot of "annotated" unimodal data, as we need to ensure that each multimodal pair is related in a meaningful way. The situation is worse when we consider more complicated multimodal settings such as language–vision, where one-to-one or one-to-many correspondence between instances of the two datasets are required, due to the difficulty in categorising data such that commonality amongst samples is preserved within categories. See Figure 1 for an example from the CUB dataset (Welinder et al., 2010a); although the same species of bird is featured in both image-caption pairs, their content differs considerably. It would be unreasonable to apply the caption from one to describe the bird depicted in the other, necessitating one-to-one correspondence between images and captions.



(a) Previous (b) Ours

Figure 2: Graphical models for multimodal generative process.

However, the scope of multimodal learning has been limited to leveraging the commonality between "related" pairs, while largely ignoring "unrelated" samples potentially available in any multimodal dataset—constructed through random pairing between modalities (Figure 3). We posit that if a distinction can be established between the "related" and "unrelated" observations within a multimodal dataset, we could greatly reduce the amount of related data required for effective learning. Figure 2 formalises this proposal. Multimodal generative models in previous work (Figure 2a) typically assumes one latent variable $\boldsymbol{z}$ that *always* generates related multimodal pair $(\boldsymbol{x}, \boldsymbol{y})$. In this work (Figure 2b), we introduce an additional Bernoulli random variable r that dictates the "relatedness" between $\boldsymbol{x}$ and $\boldsymbol{y}$ through $\boldsymbol{z}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are related when $r = 1$, and unrelated when $r = 0$.

While r can encode different dependencies, here we make the simplifying assumption that the Pointwise Mutual Information (PMI) between $\boldsymbol{x}$ and $\boldsymbol{y}$ should be high when $r = 1$, and low when $r = 0$. Intuitively, this can be achieved by adopting a max-margin metric. We therefore propose to train the generative moels with a novel contrastive-style loss (Hadsell et al., 2006; Weinberger and Saul, 2009), and demonstrate the effectiveness of our proposed method from a few different perspectives: **Improved multimodal learning:** showing improved multimodal learning for various state-of-the-art multimodal generative models on two challenging multimodal datasets. This is evaluated on four different metrics (Shi et al., 2019) summarised in § 4.2; **Data efficiency:** learning generative models under the contrastive framework requires only 20% of the data needed in baseline methods to achieve similar performance—holding true across different models, datasets and metrics; **Label propagation:** the contrastive loss encourages a larger discrepency between related and unrelated data, making it possible to directly identify related samples using the PMI between observations. We show that these data pairs can be used to further improve the learning of the generative model.

2 Related Work

Contrastive loss Our work aims to encourage data-efficient multimodal generative-model learning using a popular *representation learning* metric—contrastive loss (Hadsell et al., 2006; Weinberger and Saul, 2009). Although learning robust generalisable representations have been an important desideratum in generative-model learning (Tenenbaum and Freeman, 2000; Bengio et al., 2013), most applications of contrastive methods skip the generative component, and directly learn representations for specific tasks, as seen in contrastive predictive coding for time series data (van den Oord et al., 2018), image classification (Hénaff et al., 2019), noise contrastive estimation for vector embeddings of words (Mnih and Kavukcuoglu, 2013), as well as a range of frameworks such as DIM (Hjelm et al., 2019), MoCo (He et al., 2019), SimCLR (Chen et al., 2020) for more general visual-representation learning. These approaches, while effective for the particular task at hand, fall short in learning representations that are generalisable to different tasks.

Contrastive methods have also been employed under a generative-model setting, but typically on generative adversarial networks (GANs) to either preserve or identify factors-of-variations in their inputs. For instance, SiGAN (Hsu et al., 2019) uses a contrastive loss to preserve identity for face-image hallucination from low-resolution photos, while (Yildirim et al., 2018) uses a contrastive loss to disentangle the factors of variations in the latent code of GANs. We here employ a contrastive loss in a distinct setting of multimodal generative model learning, that, as we will show with our experiments and analyses promotes better, more robust representation learning.

Multimodal VAEs We also demonstrate that our approach is applicable across different approaches to learning multimodal generative models. To do so, we first summarise past work on multimodal VAE into two categories based on the modelling choice of approximated posterior $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$:

a) **Explicit joint models** where q_Φ is explicitly modelled as one joint encoder $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$.
Example work in this area include JMVAE (Suzuki et al., 2017), triple ELBO (Vedantam et al., 2018) and MFM (Tsai et al., 2019). Since the joint encoder require multimodal pair $(\boldsymbol{x}, \boldsymbol{y})$ as input, these approaches typically require additional modelling components and/or inference steps to deal with missing modality at test time; in fact, all three approaches propose to train unimodal VAEs on top of the joint model that handles data from each modality independently.

b) **Factorised joint models** where q_Φ is factorised as some combination f of two unimodal encoders, i.e. $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) = f\left(q_{\phi_x}(\boldsymbol{z}|\boldsymbol{x}), q_{\phi_y}(\boldsymbol{z}|\boldsymbol{y})\right)$.
This was first seen in Wu and Goodman (2018), as the MVAE model with f defined as a product of experts (PoE), i.e. $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) = q_{\phi_x}(\boldsymbol{z}|\boldsymbol{x}) q_{\phi_y}(\boldsymbol{z}|\boldsymbol{y}) p(\boldsymbol{z})$, allowing for cross-modality generation without extra modelling components. Particularly, the MVAE was constructed to cater to multimodal settings where data was not guaranteed to be organised into related sets, and where additional modalities were taken to be, in terms of information content, subsets of a primary data source—such as images and their class labels.
Alternately, Shi et al. (2019) explored an approach that explicitly leveraged the availability of related/paired data, motivated by arguments from embodied cognition of the world. They propose the MMVAE model, which additionally differs from the MVAE model in its choice of posterior approximation—where f is modelled as the mixture of experts (MoE) of unimodal posteriors—to ameliorate shortcomings to do with precision miscalibration of the PoE. Furthermore, Shi et al. (2019) also posit four criteria that a multimodal VAE should satisfy, which we adopt in this work to evaluate the performance of our models.

3 METHODOLOGY

Given data over observations from two modalties $(\boldsymbol{x}, \boldsymbol{y})$, one can learn a multimodal VAE targetting $p_\Theta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{z}) p_{\theta_x}(\boldsymbol{x}|\boldsymbol{z}) p_{\theta_y}(\boldsymbol{y}|\boldsymbol{z})$, where $p_\theta(\cdot|\boldsymbol{z})$ are deep neural networks (decoders) parametrised by $\Theta = \{\theta_x, \theta_y\}$. To maximise the joint marginal likelihood $\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})$, one approximates the intractable model posterior $p_\Theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ with a variational posterior $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$, allowing us to optimise a variational evidence lower bound (ELBO), defined as

$$\log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) \geq \mathbb{E}_{\boldsymbol{z} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})} \left[\log \frac{p_\Theta(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y})}{q_\Phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})} \right] = \text{ELBO}(\boldsymbol{x}, \boldsymbol{y}). \tag{1}$$

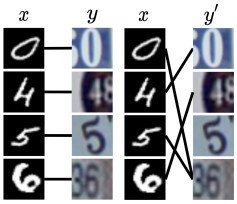
This leaves open the question of how to model the approximatd posterior $q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$. As mentioned in § 2, there are two schools of thinking, namely *explicit joint model* such as JMVAE (Suzuki et al., 2017) and *factorised joint model* including MVAE (Wu and Goodman, 2018) and MMVAE (Shi et al., 2019). In this work we demonstrate the effectiveness of our approach across all these models.

3.1 CONTRASTIVE LOSS FOR "RELATEDNESS" LEARNING

Where prior work always assumes $(\boldsymbol{x}, \boldsymbol{y})$ to be related, we introduce a *relatedness* variable r explicitly capturing this aspect. Our approach is motivated by the characteristics of pointwise mutual information (PMI) between related and unrelated observations across modalities:

Hypothesis 3.1. *Let $(\boldsymbol{x}, \boldsymbol{y}) \sim p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ be a related data pair from two modalities, and let $\boldsymbol{y}'$ denote a data point unrelated to $\boldsymbol{x}$. Then, the pointwise mutual information $I(\boldsymbol{x}, \boldsymbol{y}) > I(\boldsymbol{x}, \boldsymbol{y}')$.*

Note that the PMI measures the statistical dependence between values (x, y), which for a joint distribution $p(x, y)$ is defined as $I(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$. Hypothesis 3.1 should be a fairly uncontroversial assumption: we say simply that under the joint distribution for related data $p_\Theta(\boldsymbol{x}, \boldsymbol{y})$, the PMI between related points $(\boldsymbol{x}, \boldsymbol{y})$ is stronger than that between unrelated points $(\boldsymbol{x}, \boldsymbol{y}')$. It is therefore possible to leverage relatedness by maximising the difference in PMI between related and unrelated pairs, i.e. $I(\boldsymbol{x}, \boldsymbol{y}) - I(\boldsymbol{x}, \boldsymbol{y}')$. We show in Appendix A that this is equivalent to maximising the difference between the joint marginal likelihoods of related and unrelated pairs (Figure 3a and 3b). This casts multimodal learning as max-margin optimisation, with the contrastive (triplet) loss as a natural choice of objective: $\mathcal{L}_C(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}') = d(\boldsymbol{x}, \boldsymbol{y}) - d(\boldsymbol{x}, \boldsymbol{y}') + m$. We adopt this loss to our objective by replacing d with the (negative) joint marginal likelihood $-\log p_\Theta$. Following Song et al. (2016), with N negative samples $\{y'_i\}_{i=1}^N$, we have



(a) Related (b) Unrelated

Figure 3: Constructing related & unrelated samples

$$\mathcal{L}_C(\boldsymbol{x}, Y) = -\log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) + \log \sum_{i=1}^{N} p_\Theta(\boldsymbol{x}, \boldsymbol{y}'_i). \tag{2}$$

As the loss is asymmetric, one can average over $\mathcal{L}_C(\boldsymbol{y}, X)$ and $\mathcal{L}_C(\boldsymbol{x}, Y)$ to ensure negative samples in both modalities are accounted for—giving our contrastive objective:

$$\mathcal{L}_C(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\{\mathcal{L}_C(\boldsymbol{x}, Y) + \mathcal{L}_C(\boldsymbol{y}, X)\} = \underbrace{-\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})}_{①} + \underbrace{\frac{1}{2}\left(\log \sum_{\boldsymbol{x}'_i=1}^{N} p_\Theta(\boldsymbol{x}', \boldsymbol{y}) + \log \sum_{\boldsymbol{y}'_i=1}^{N} p_\Theta(\boldsymbol{x}, \boldsymbol{y}') \right)}_{②} \tag{3}$$

Note that since only the joint marginal likelihood terms are needed in (3), the contrastive learning framework can be directly applied to *any* multimodal generative model without needing extra components. We also show our contrastive framework applies when the number of modalities is more than two (cf. Appendix C).

Dissecting $\mathcal{L}_C$ Although similar to the VAE (1), our objective (3) directly maximises $p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ in ①. $\mathcal{L}_C$ by itself is not a effective objective as ② in $\mathcal{L}_C$ *minimises* $p_\Theta(\boldsymbol{x}, \boldsymbol{y})$, which can overpower the effect of ① during training.

Figure 4: $\log p(\boldsymbol{x}, \boldsymbol{y})$ of imitation, unrelated digits and random noise, where $p = \mathcal{N}(\boldsymbol{m_x}, \boldsymbol{m_y})$.

We intuit this phenomenon using a simple example in Figure 4, showing the log likelihood of $\log p(\boldsymbol{x}, \boldsymbol{y})$ in column 2, 3, 4 (green) on a Gaussian distribution $\mathcal{N}((\boldsymbol{m_x}, \boldsymbol{m_y}); c)$, with the images in the first

column of Figure 4 as means and with constant variance c. While achieving high log likelihood $\log p(\boldsymbol{x}, \boldsymbol{y})$ on this distribution requires matching $(\boldsymbol{m}_{\boldsymbol{x}}, \boldsymbol{m}_{\boldsymbol{y}})$ (col 2), we see both unrelated digits (col 3) and noise (col 4) can lead to (comparatively) poor joint log likelihoods. This indicates that the generative model need not generate valid, unrelated images to minimise ②—generating noise would have roughly the same effect on log likelihood. As a result, the model can trivially minimise $\mathcal{L}_C$ by generating noise that minimises ② instead of accurate reconstruction that maximises ①.

This learning dynamic is verified empirically, as we show in Figure 9 in Appendix D: optimising $\mathcal{L}_C$ by itself results in the loss approaching 0 within the first 100 iterations, and both ① and ② takes on extremely low values, resulting in a model that generates random noise.

Final objective To mitigate this issue, we need to ensure that minimising ② does not overpower maximising ①. We hence introduce a hyperparameter γ on ① to upweight the maximisation of the marginal likelihood, with the final objective to minimise

$$\mathcal{L}_C(\boldsymbol{x}, \boldsymbol{y}) = -\gamma \underbrace{\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})}_{①} + \underbrace{\frac{1}{2}\left(\log \sum_{\boldsymbol{x}'_i=1}^{N} p_\Theta(\boldsymbol{x}', \boldsymbol{y}) + \log \sum_{\boldsymbol{y}'_i=1}^{N} p_\Theta(\boldsymbol{x}, \boldsymbol{y}')\right)}_{②}, \qquad \gamma > 1. \quad (4)$$

We conduct ablation studies on the effect of γ in Appendix E, noting that larger γ encourages better quality of generation and more stable training in some cases, while models trained with smaller γ are better at predicting "relatedness" between multimodal samples. We also note that optimising (4) maximises the pointwise mutual information $I(\boldsymbol{x}, \boldsymbol{y})$; see Appendix B for a proof.

3.2 Optimising the objective

Since in VAEs we do not have access to the exact marginal likelihood, we have to optimise an approximated version of the true contrastive objective in (4). In (5) we list a few possible candidates of estimators and their relationships to the (log) joint marginal likelihood:

$$\text{ELBO} \leq \underbrace{\mathbb{E}_{\{\boldsymbol{z}_k\}_1^K \sim q_\Phi}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\Theta(\boldsymbol{z}_k, \boldsymbol{x}, \boldsymbol{y})}{q_\Phi(\boldsymbol{z}_k \mid \boldsymbol{x}, \boldsymbol{y})}\right]}_{\text{IWAE}} \leq \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) \leq \underbrace{\mathbb{E}_{\{\boldsymbol{z}_k\}_1^K \sim q_\Phi}\left[\log \sqrt[2]{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{p_\Theta(\boldsymbol{z}_k, \boldsymbol{x}, \boldsymbol{y})}{q_\Phi(\boldsymbol{z}_k \mid \boldsymbol{x}, \boldsymbol{y})}\right)^2}\right]}_{\text{CUBO}} \quad (5)$$

with the ELBO from Eq. (1), the importance weighted autoencoder (IWAE), a K-sample lower bound estimator that can compute an arbitrarily tight bound with increasing K (Burda et al., 2015), and the χ upper bound (CUBO), an *upper-bound* estimator (Dieng et al., 2017).

We now discuss our choice of approximation of $\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ for each term in equation (4):

① Minimising this term maximises the joint marginal likelihood, which can hence be approximated with a valid lower-bound estimator (5); the IWAE estimator being the preferred choice.

② Since this term *minimises* the joint marginal likelihoods, employing an upper-bound estimator (CUBO in (5)) will ensure an upper-bound estimator to $\mathcal{L}_C$ (4). While such a bounded approximation is indeed desirable, existing upper-bound estimators tend to have rather large bias/variance and can therefore yield poor quality approximations.

We therefore also estimate ② with the IWAE estimator, as it provides an *arbitrarily tight, low variance lower-bound* (Burda et al., 2015) to $\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})$, in the hope that having a more accurate approximation to the marginal likelihood (and by extension, the contrastive loss) can better affect the performance of the model, in contrast to ensuring a valid bound on the objective. We report results using both IWAE and CUBO estimators in our experiment, denoted as `cI` and `cC` respectively.

4 Experiments

As stated in § 1, we analyse the suitability of contrastive learning for multimodal generative models from three persepctives—*improved multimodal learning* (§ 4.3), *data efficiency* (§ 4.4) and *label propagation* (§ 4.5). Throughout the experiments, we take $N = 5$ negative samples for the contrastive objective, set $\gamma = 2$ based on analyses of ablations in appendix E, and take $K = 30$ samples for our IWAE estimators. We now introduce the datasets and metrics used for our experiments.

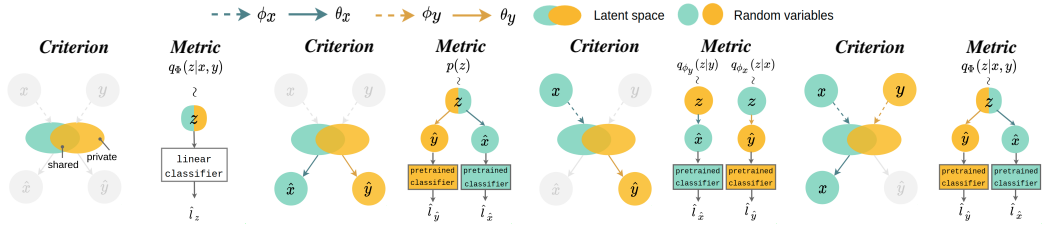
4.1 DATASETS

MNIST-SVHN The dataset is designed to separate conceptual complexity, i.e. digit, from perceptual complexity, i.e. color, style, size. Each data pair contains 2 samples of the same digit, one from each dataset (see examples in Figure 3a). We construct the dataset such that each instance from one dataset is paired with 30 instances of the same digit from the other dataset. Although both datasets are simple and well-studied, the many-to-many pairing between samples creates matching of different writing styles vs. backgrounds and colors, making it a challenging multimodal dataset.

CUB Image-Captions We also consider a more challenging language-vision multimodal dataset, Caltech-UCSD Birds (CUB) (Welinder et al., 2010b; Reed et al., 2016). The dataset contains 11,788 photos of birds, paired with 10 captions describing the bird's physical characteristics, collected through Amazon Mechanical Turk (AMT). See CUB image-caption pair in Figure 1.

4.2 METRICS

Shi et al. (2019) proposed four criteria for multimodal generative models (Figure 5, left), that we summarise and unify as metrics to evaluate these criteria for different generative models (Figure 5, right). We now introduce each criterion and its corresponding metric in detail.



(a) Latent factorisation (b) Coherent joint generation (c) Coherent cross generation (d) Synergy

Figure 5: ***Left of each pair:*** Four criteria for multi-modal generative models; image adapted from Shi et al. (2019). ***Right of each pair:*** Four metrics to evaluate the model's performance on criterion in corresponding row.

(a) Latent accuracy (Figure 5a) *Criterion: latent space factors into "shared" and "private" subspaces across modalities.* We fit a linear classifier on the samples from $\boldsymbol{z} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ to classify the information shared between the two modalities. For MNIST-SVHN, this can be the digit label as shown in Figure 5a (right). We check if $\hat{l}_z$ is the same as the digit label of the original inputs $\boldsymbol{x}$ and $\boldsymbol{y}$, with the intuition that extracting the commonality between $\boldsymbol{x}$ and $\boldsymbol{y}$ from latent representation using a linear transform, supports the claim that the latent space has factored as desired.

(b) Joint coherence (Figure 5b) *Criterion: model generates paired samples that preserves the commonality observed in data.* Again taking MNIST-SVHN as an example, this can be verified by taking pre-trained MNIST and SVHN digit classifiers and applying them on the multimodal observations generated from the same prior sample z. Coherence is computed by how often generations $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ classify to the same digit, i.e. whether $\hat{l}_{\hat{x}} = \hat{l}_{\hat{y}}$.

(c) Cross coherence (Figure 5c) *Criterion: model generates data in one modality conditioned on the other while preserving shared commonality.* To compute cross coherence, we generate observations $\hat{\boldsymbol{x}}$ using latent from unimodal marginal posterior $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{y})$ and $\hat{\boldsymbol{y}}$ from $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$. Similar to joint coherence, the criterion here is evaluated by predicting the label of the cross-generated samples $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$ using off-the-shelf MNIST and SVHN classifiers. In Figure 5c (right), the cross coherence is the frequency of which $\hat{l}_{\hat{\boldsymbol{x}}} = l_{\boldsymbol{y}}$ and $\hat{l}_{\hat{\boldsymbol{y}}} = l_{\boldsymbol{x}}$.

(d) Synergy coherence (Figure 5d) *Criterion: models learnt across multiple modalities should be no worse than those learnt from just one.* For consistency, we evaluate this criterion from a coherence perspective. Given generations $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ from $\boldsymbol{z} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$, we again examine if generated and original labels match; i.e. if $\hat{l}_{\hat{\boldsymbol{x}}} = l_{\boldsymbol{y}} = \hat{l}_{\hat{\boldsymbol{y}}} = l_{\boldsymbol{x}}$.

See Appendix F for details on architecture. All quantitative results are reported over 5 runs.

4.3 IMPROVED MULTIMODAL LEARNING

Finding: *Contrastive learning improves multimodal learning across all models and datasets.*

MNIST-SVHN See Table 1 (*top, 100% data used*) for results on the full MNIST-SVHN dataset. Note that for MMVAE, since the joint posterior q_Φ is factorised as the mixture of unimodal posteriors q_{ϕ_x} and q_{ϕ_y}, the model never directly takes sample from the explicit form of the joint posterior. Instead, it takes equal number of samples from each unimodal posteriors, reflective of the equal

weighting of the mixture. As a result, it is not meaningful to compute synergy coherence for MMVAE as it is exactly the same as the coherence of any single-way generation.

From Table 1 (*top, 100% data used*), we see that models trained on our contrastive objective (cI-<MODEL> and cC-<MODEL>) improves multimodal learning performance significantly for all three generative models evaluated on the metrics. The results showcase the robustness of our approach from the perspectives of modelling choice and metric of interests. In particular, note that for the best performing model MMVAE, using IWAE estimator for ② of Eq (4) (cI-MMVAE) yields slightly better results than CUBO (cC-MMVAE), while for the two other models the performance for different estimators are similar.

Table 1: Evaluation of baselines MMVAE, MVAE, JMVAE and their contrastive variations (cI-<MODEL>, cC-<MODEL> for the IWAE and CUBO estimators used in Eq (4), respectively), on MNIST(M)-SVHN(S) dataset, using 100% (top) and 20% (bottom) of data.

Data	Methods	Latent accuracy (%)		Joint Coherence (%)	Cross coherence (%)		Synergy coherence (%)	
		M	S		S→M	M→S	joint→M	joint→S
100% of data used	MMVAE	92.48 (±0.37)	79.03 (±1.17)	42.32 (±2.97)	70.77 (±0.35)	85.50 (±1.05)	—	—
	cI-MMVAE	**93.97 (±0.36)**	**81.87 (±0.52)**	43.94 (±0.96)	**79.66 (±0.59)**	**92.67 (±0.29)**	—	—
	cC-MMVAE	93.10 (±0.17)	80.88 (±0.80)	**45.46 (±0.78)**	79.34 (±0.54)	92.35 (±0.46)	—	—
	MVAE	91.65 (±0.17)	64.12 (±4.58)	9.42 (±7.82)	10.98 (±0.56)	21.88 (±2.21)	64.60 (±9.25)	52.91 (±8.11)
	cI-MVAE	96.97 (±0.84)	75.94 (±6.20)	15.23 (±10.46)	10.85 (±1.17)	27.70 (±2.09)	85.07 (±7.73)	75.67 (±4.13)
	cC-MVAE	97.42 (±0.40)	81.07 (±2.03)	8.85 (±3.86)	12.83 (±2.25)	30.03 (±2.46)	75.25 (±5.31)	69.42 (±3.94)
	JMVAE	84.45 (±0.87)	57.98 (±1.27)	42.18 (±1.50)	49.63 (±1.78)	54.98 (±3.02)	85.77 (±0.66)	68.15 (±1.38)
	cI-JMVAE	84.58 (±1.49)	64.42 (±1.42)	48.95 (±2.31)	58.16 (±1.83)	70.61 (±3.13)	**93.45 (±0.52)**	84.00 (±0.97)
	cC-JMVAE	83.67 (±3.48)	66.64 (±2.92)	47.27 (±4.52)	59.73 (±3.85)	69.49 (±2.19)	91.21 (±6.59)	**84.07 (±3.19)**
20% of data used	MMVAE	88.54 (±0.37)	68.90 (±1.79)	37.71 (±0.60)	59.52 (±0.28)	76.33 (±2.23)	—	—
	cI-MMVAE	91.64 (±0.06)	**73.02 (±0.80)**	**42.74 (±0.36)**	**69.51 (±1.18)**	**86.75 (±0.28)**	—	—
	cC-MMVAE	92.10 (±0.19)	71.29 (±1.05)	40.77 (±0.93)	68.43 (±0.90)	86.24 (±0.89)	—	—
	MVAE	90.29 (±0.57)	33.44 (±0.26)	10.88 (±9.15)	8.72 (±0.92)	12.12 (±3.38)	42.10 (±5.22)	44.95 (±5.92)
	cI-MVAE	**93.72 (±1.09)**	56.74 (±7.97)	12.79 (±6.82)	14.18 (±2.19)	20.23 (±4.55)	75.36 (±5.05)	64.81 (±4.81)
	cC-MVAE	92.74 (±2.97)	52.99 (±8.35)	17.95 (±12.52)	14.70 (±1.65)	24.90 (±5.77)	56.86 (±18.84)	54.28 (±9.86)
	JMVAE	77.53 (±0.13)	52.55 (±2.18)	26.37 (±0.54)	42.58 (±5.32)	41.44 (±2.26)	85.07 (±9.74)	51.95 (±2.28)
	cI-JMVAE	77.57 (±4.02)	57.91 (±1.28)	32.58 (±5.89)	51.85 (±1.27)	47.92 (±10.32)	**92.54** (±1.13)	67.01 (±8.72)
	cC-JMVAE	81.11 (±2.76)	57.85 (±2.23)	34.00 (±7.18)	50.73 (±0.45)	56.89 (±6.18)	88.36 (±4.38)	**68.49** (±8.82)

CUB Following Shi et al. (2019), for the images in CUB, we observe and generate in feature space instead of pixel space by preprocessing the images using a pre-trained ResNet-101 (He et al., 2016). A nearest-neighbour lookup among all the features in the test set is used to project the feature generations of the model back to image space. This helps circumvent CUB image complexities to some extent—as the primary goal here is to learn good models and representations of multimodal data, rather than a focus on pixel-level image quality of generations.

The metrics listed in § 4.2 can also be applied to CUB with some modifications. Since bird-species classes are disjoint for the train and test sets, and as we show in Figure 1 contains substantial in-class variance, it is not constructive to evaluate these metrics using bird categories as labels. In Shi et al. (2019), the authors propose to use Canonical Correlation Analysis (CCA)—used by Massiceti et al. (2018) as a reliable vision-language correlation baseline—to compute coherence scores between generated image-caption pairs; which we employ (i.e. (b), (c), (d) in Figure 5) for CUB.

We show the results in Table 2 (*top, 100% data used*). We see that our contrastive approach (both cI-<MODEL> and cC-<MODEL>) is even more effective on this challenging vision-language dataset, with significant improvements to the correlation of generated image-caption pairs. It is also on this more complicated dataset where the advantages of using the stable, low-variance IWAE estimator are highlighted — for both MVAE and JMVAE, the contrastive objective with CUBO estimator suffers from numerical stability issues, yielding close-to-zero correlations for all metrics in Table 2. Results for these models are therefore omitted.

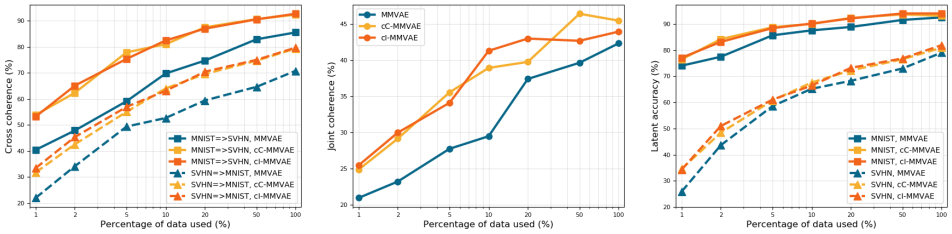
4.4 Data Efficiency

Finding: *Contrastive learning on 20% of data matches baseline models on full data.*

We plot the quantitative performance of MMVAE with and without contrastive learning against the percentage of the original dataset used, as seen in Figure 6. We observe that the performance of contrastive MMVAE with the IWAE (*cI-MMVAE, red*) and CUBO estimators (*cC-MMVAE, yellow*) are *consistently* better than the baselines (*MMVAE, blue*), and that baseline performance using *all* related data is matched by the contrastive MMVAE using just 10—20% of data. The partial datasets used

Table 2: Evaluation of baseline MMVAE, MVAE, JMVAE and their contrastive variations (cI-<MODEL> and cC-<MODEL>) on CUB image(img)-caption(cap) dataset, using 100% (top) and 20% (bottom) of data.

Data	Methods	Joint Coherence (CCA)	Cross coherence (CCA)		Synergy coherence (CCA)	
			img→cap	cap→ img	joint→img	joint→cap
100% of data used	MMVAE	0.212 (±*2.94e-2*)	0.154 (±*7.05e-3*)	0.244 (±*5.83e-3*)	-	-
	cI-MMVAE	**0.314** (±***3.12e-2***)	0.188 (±*4.02e-3*)	0.334 (±*1.20e-2*)	-	-
	cC-MMVAE	0.263 (±*1.47e-2*)	0.244 (±*2.24e-2*)	**0.369** (±***3.18e-3***)	-	-
	MVAE	0.075 (±*8.71e-2*)	-0.008 (±*1.41e-4*)	0.000 (±*1.84e-3*)	-0.002 (±*4.95e-3*)	0.001 (±*1.13e-3*)
	cI-MVAE	0.209 (±*3.12e-2*)	0.247 (±*1.12e-4*)	-0.008 (±*3.99e-4*)	**0.218** (±*5.27e-3*)	0.148 (±*9.23e-3*)
	JMVAE	0.220 (±*1.19e-2*)	0.157 (±*4.98e-2*)	0.191 (±*3.11e-2*)	0.212 (±*1.23e-2*)	0.143 (±*1.14e-1*)
	cI-JMVAE	0.255 (±*3.24e-3*)	0.149 (±*1.25e-3*)	0.226 (±*7.48e-2*)	0.202 (±*1.12e-4*)	**0.176** (±***6.23e-2***)
20% of data used	MMVAE	0.117 (±*1.51e-2*)	0.094 (±*7.21e-3*)	0.153 (±*1.47e-2*)	-	-
	cI-MMVAE	0.206 (±*1.65e-2*)	0.136 (±*1.51e-2*)	0.251 (±*2.39e-2*)	-	-
	cC-MMVAE	0.226 (±*4.69e-2*)	0.188 (±*2.80e-2*)	**0.273** (±***8.67e-3***)	-	-
	MVAE	0.091 (±*2.63e-2*)	0.008 (±*4.81e-3*)	0.005 (±*7.50e-3*)	0.020 (±*8.77e-3*)	0.009 (±*1.17e-2*)
	cI-MVAE	0.132 (±*3.33e-2*)	**0.192** (±***3.91e-2***)	-0.002 (±*3.61e-3*)	0.162 (±*7.89e-2*)	0.081 (±*3.82e-2*)
	JMVAE	0.127 (±*3.76e-2*)	0.118 (±*3.82e-3*)	0.154 (±*8.34e-3*)	0.181 (±*1.26e-2*)	0.139 (±*1.33e-2*)
	cI-JMVAE	**0.269** (±***1.20e-2***)	0.134 (±*4.24e-4*)	0.210 (±*2.35e-2*)	**0.192** (±***1.41e-4***)	**0.168** (±***3.82e-3***)



(a) Cross coherence (b) Joint coherence (c) Latent accuracy

Figure 6: Performance of **MMVAE**, **cI-MMVAE** and **cC-MMVAE** using $n\%$ of MNIST-SVHN.

here are constructed by first taking $n\%$ of each unimodal dataset, then pairing to create multimodal datasets (§ 4.1)—ensuring it contains the requisite amount of "related" samples. In addition, we reproduce results generated from using 100% of the data in MNIST-SVHN and CUB (Tables 1 and 2) using only 20% of the original multimodal datasets (Tables 1 and 2 (*bottom*)). Comparing results between *top* vs. *bottom* in Tables 1 and 2 shows that this finding holds across the models, on both MNIST-SVHN and CUB datasets. This shows that the efficiency gains from a contrastive approach is invariant to VAE type, data, and metrics used, underscoring its effectiveness.

4.5 Label Propagation

Finding: *Generative models learned contrastively are good predictors of "relatedness", enabling label propagation and matching baseline performance on full datasets, using only 10% of data.*

Here, we show that our contrastive framework encourages a larger discrepancy between the PMI of related vs. unrelated data, as set out in hypothesis 3.1, allowing one to first train the model on a small subset of related data, and subsequently construct a classifier using PMI that identifies related samples in the remaining data. We now introduce our pipeline for label propagation in details.

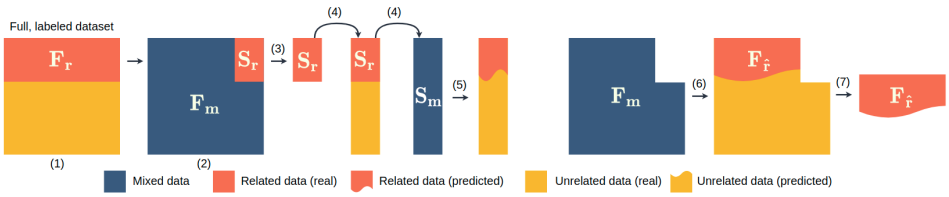
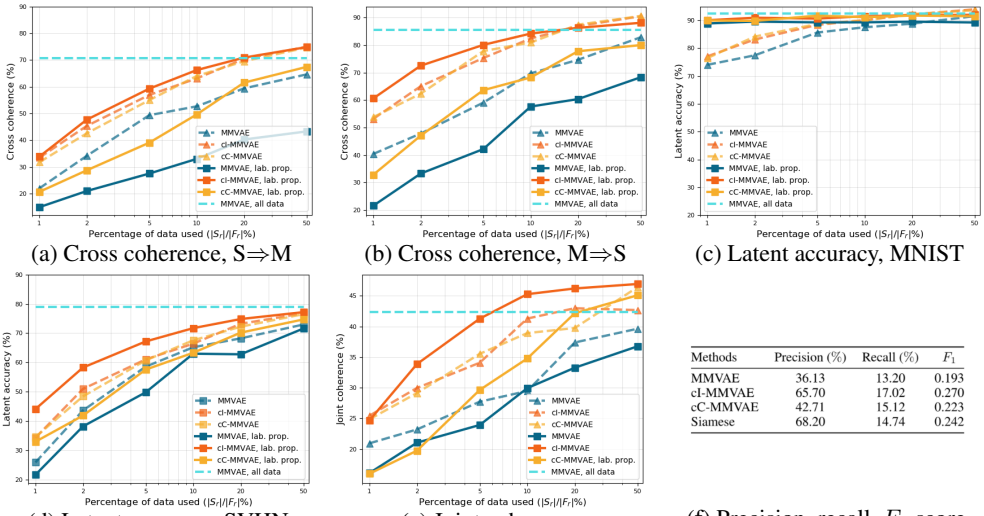


Figure 7: Pipeline of label propagation

Pipeline As showing in Figure 7, we first construct a full dataset by randomly matching instances in MNIST and SVHN, and denote the related pairs by F_r (full, related). We further assume access to only $n\%$ of F_r, denoted as S_r (small, related), and denote the rest as F_m, containing a mix of related and unrelated pairs. Next, we train a generative model g on S_r. To find a relatedness threshold, we construct a small, mixed dataset S_m by randomly matching samples across modalities in S_r. Given relatedness ground-truth for S_m, we can compute the PMI $I(\boldsymbol{x}, \boldsymbol{y}) = \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) - \log p_{\theta_x}(\boldsymbol{x}) p_{\theta_y}(\boldsymbol{y})$ for all pairs $(\boldsymbol{x}, \boldsymbol{y})$ in S_m and estimate an optimal threshold. This threshold can



(a) Cross coherence, S⇒M (b) Cross coherence, M⇒S (c) Latent accuracy, MNIST

(d) Latent accuracy, SVHN (e) Joint coherence

Methods	Precision (%)	Recall (%)	F_1
MMVAE	36.13	13.20	0.193
cI-MMVAE	65.70	17.02	0.270
cC-MMVAE	42.71	15.12	0.223
Siamese	68.20	14.74	0.242

(f) Precision, recall, F_1 score

Figure 8: Models with and without label propagation using **MMVAE**, **cI-MMVAE** and **cC-MMVAE**.

now be applied to the full, mixed dataset F_m to identify related pairs giving us a new related dataset $F_{\hat{r}}$, which can be used to further improve the performance of the generative model g.

Results In Figure 8 (a-e), we plot the performance of baseline MMVAE (blue) and contrastive MMVAE with IWAE (cI-MMVAE, red) and CUBO (cC-MMVAE, yellow) estimators for term ② of Eq (4), trained with (solid curves) and without (dotted curves) label propagation. Here, the x-axis is the proportion in size of S_r to F_r, i.e. the percentage of related data used to pretrain the generative model before label propagation. We compare these results to MMVAE trained on all related data F_r (cyan horizontal line) as a "best case scenario" of these training regimes.

Clearly, label propagation using a contrastive model with IWAE estimator is helpful, and in general the improvement is greater when less data is available; Figure 8 also shows that when S_r is 10% of F_r, cI-MMVAE is competitive with the performance of baseline MMVAE trained on F_r.

For baselines MMVAE and cC-MMVAE however, label propagation hurts performance no matter the size of S_r, as shown by the blue and yellow curves in Figure 8 (a-e). The advantages of cI-MMVAE is further demonstrated in Figure 8f, where we compute the precision, recall, and F_1 score of relatedness prediction on F_m, for models trained on 10% of all related data. We also compare to a simple label-propagation baseline, where the relatedness of F_m is predicted using a siamese network (Hadsell et al., 2006) trained on the same 10% dataset. Notably, while the Siamese baseline in Figure 8f is a competitive predictor of relatedness, cI-MMVAE has the highest F_1 score amongst all four, making it the most reliable indicator of relatedness. Beyond that, note that with the contrastive MMVAE, relatedness can be predicted *without additional training* and only requires a simple threshold computation directly computed using the generative model. The fact that the cI-MMVAE's relatedness-prediction performance is the *only* one that matches the Siamese baseline strongly supports the view that the contrastive loss encourages generative models to utilise and better learn the relatedness between multimodal pairs; in addition, the poor performance of cC-MMVAE shows that the advantages of having a bounded estimation to the contrastive objective by using an upper-bound for ② is overshadowed by the high bias of CUBO, and that one may benefit more from choosing a low variance lower-bound like IWAE.

5 Conclusion

We introduced a contrastive-style objective for multimodal VAE, aiming at reducing the amount of multimodal data needed by exploiting the distinction between "related" and "unrelated" multimodal pairs. We showed that this objective improves multimodal training, drastically reduce the amount of multimodal data needed, and establishes a strong sense of "relatedness" for the generative model. These findings hold true across a multitude of datasets, models and metrics. The positive results of our method indicates that it is beneficial to utilise the relatedness information when training on multimodal data, which has been largely ignored in previous work. While we propose to utilise it implicitly through contrastive loss, future work may consider relatedness as a random variable in the graphical model and see if explicit dependency on relatedness can be useful.

REFERENCES

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

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2015.

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

A. B. Dieng, D. Tran, R. Ranganath, J. W. Paisley, and D. M. Blei. Variational inference via χ upper bound minimization. In *31st Annual Conference on Neural Information Processing Systems, NIPS 2017*, pages 2732–2741, 2017.

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

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

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

K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019.

O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

D. Massiceti, P. K. Dokania, N. Siddharth, and P. H. Torr. Visual dialogue without vision or dialogue. In *NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*, 2018.

A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pages 2265–2273, 2013.

S. E. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. *CoRR*, abs/1605.05395, 2016.

Y. Shi, S. Narayanaswamy, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 15718–15729, 2019.

H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016.

M. Suzuki, K. Nakayama, and Y. Matsuo. Joint multimodal learning with deep generative models. In *International Conference on Learning Representations Workshop*, 2017.

J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

Y. H. Tsai, P. P. Liang, A. A. Bagherzade, L.-P. Morency, and R. Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

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

R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations*, 2018.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010a.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010b.

M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5580–5590, 2018.

G. Yildirim, N. Jetchev, and U. Bergmann. Unlabeled disentangling of gans with guided siamese networks. 2018.

I. Yildirim. From perception to conception: Learning multisensory representations. 2014.

C. Zhang, Z. Yang, X. He, and L. Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1, 2020.

Appendix:

A FROM POINTWISE MUTUAL INFORMATION TO JOINT MARGINAL LIKELIHOOD

In this section, we show that for the purpose of utilising the relatedness between mutlimodal pairs, maximising the difference between *pointwise mutual information* between related points $\boldsymbol{x}, \boldsymbol{y}$ and unrelated points $\boldsymbol{x}, \boldsymbol{y}'$ is equivalent to maximising the difference between their log *joint marginal likelihoods*.

To see this, we can expand $I(\boldsymbol{x}, \boldsymbol{y}) - I(\boldsymbol{x}, \boldsymbol{y}')$ as

$$\begin{aligned} & I(\boldsymbol{x}, \boldsymbol{y}) - I(\boldsymbol{x}, \boldsymbol{y}') \\ & \quad = [\log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) - \log p_\Theta(\boldsymbol{x}) - \log p_\Theta(\boldsymbol{y})] - [\log p_\Theta(\boldsymbol{x}, \boldsymbol{y}') - \log p_\Theta(\boldsymbol{x}) - \log p_\Theta(\boldsymbol{y}')] \\ & \quad = \underbrace{\log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) - \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}')}_{①} + \underbrace{\log p_\Theta(\boldsymbol{y}') - \log p_\Theta(\boldsymbol{y})}_{②}. \end{aligned} \tag{6}$$

In (6) the PMI difference is decomposed as two terms: ① the difference between joint marginal likelihoods and ② the difference between marginal likelihoods of different instances of $\boldsymbol{y}$-s. It is clear that since ② involves only one modality and only accounts for the difference in values between $\boldsymbol{y}$ and $\boldsymbol{y}'$, it is not relevant to the relatedness of $\boldsymbol{x}$, $\boldsymbol{y}$.

Therefore, for the purpose of utilising relatedness information, we only need to maximise $I(\boldsymbol{x}, \boldsymbol{y}) - I(\boldsymbol{x}, \boldsymbol{y}')$ through maximising term ①, i.e. the difference between joint marginal likelihoods.

B CONNECTION OF FINAL OBJECTIVE TO POINTWISE MUTUAL INFORMATION

Here, we show that minimising the objective in (4) maximises the PMI between $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) &= -\gamma \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2}\left(\log \sum_{\boldsymbol{x}' \in X} p_\Theta(\boldsymbol{x}', \boldsymbol{y}) + \log \sum_{\boldsymbol{y}' \in Y} p_\Theta(\boldsymbol{x}, \boldsymbol{y}')\right) \\ &= -(\gamma - \frac{1}{2}) \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{2} \log \frac{p_\Theta(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in Y} p_\Theta(\boldsymbol{x}, \boldsymbol{y}') \sum_{\boldsymbol{x}' \in X} p_\Theta(\boldsymbol{y}, \boldsymbol{x}')} \\ &\approx -(\gamma - \frac{1}{2}) \log p_\Theta(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{2} \underbrace{\log \frac{p_\Theta(\boldsymbol{x}, \boldsymbol{y})}{p_\Theta(\boldsymbol{x}) p_\Theta(\boldsymbol{y})}}_{I(\boldsymbol{x}, \boldsymbol{y})} \end{aligned} \tag{7}$$

We see in (7) that minimising $\mathcal{L}$ can be decomposed to maximising both the joint marginal likelihood $p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ and $I(\boldsymbol{x}, \boldsymbol{y})$. Note that since $\gamma > 1$, we can be sure that the joint marginal likelihood weighting $\gamma - \frac{1}{2}$ is non-negative.

C GENERALISATION TO $M > 2$ MODALITIES

In this section we show how the contrastive loss generalise to cases where number of modalities considered M is greater than 2.

Given observations from M modalities $D = \{X_1, X_2, \cdots, X_m, \cdots, X_M\}$, where X_m denotes unimodal dataset of modalit m of size N_m, i.e. $X_m = \{\boldsymbol{x}_m^{(i)}\}_{i=1}^{N_m}$. Similar to (2), we can write the assymetrical contrastive loss for any observation $\boldsymbol{x}_m^{(i)}$ from modality m, where negative samples are taken for all $(M-1)$ other modalities:

$$\mathcal{L}_C(\boldsymbol{x}_m^{(i)}, D_{\tilde{m}}) = -\log p_\Theta(\boldsymbol{x}_{1:M}^{(i)}) + \log \sum_{\substack{d=1 \\ d\neq m}}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{N} p_\Theta(\boldsymbol{x}_{1:(d-1),(d+1):M}^{(i)}, \boldsymbol{x}_d^{(j)}). \tag{8}$$

We can therefore rewrite (3) as:

$$\mathcal{L}_C(\boldsymbol{x}_{1:M}^{(i)}) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_C(\boldsymbol{x}_m^{(i)}, D_{\tilde{m}}) \tag{9}$$

$$= -\log p_\Theta(\boldsymbol{x}_{1:M}^{(i)}) + \frac{1}{M} \sum_{m=1}^{M} \left(\log \sum_{\substack{d=1 \\ d\neq m}}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{N} p_\Theta(\boldsymbol{x}_{1:(d-1),(d+1):M}^{(i)}, \boldsymbol{x}_d^{(j)}) \right), \tag{10}$$

where N is the number of negative samples, all $\log p_\Theta(\boldsymbol{x}_{1:M})$ are approximated by the following joint ELBO for M modalities:

$$\log p_\Theta(\boldsymbol{x}_{1:M}) \geq \mathbb{E}_{\boldsymbol{z} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \left[\log \frac{p_\Theta(\boldsymbol{z}, \boldsymbol{x}_{1:M})}{q_\Phi(\boldsymbol{z} \mid \boldsymbol{x}_{1:M})} \right] = \text{ELBO}(\boldsymbol{x}_{1:M}). \tag{11}$$

While the above gives us the true generalisation of (3), we note that the number of times where ELBO needs to be evaluated in (10) is $\mathcal{O}(M^2 N)$, making it difficult to implement this objective in practice, especially on more complicated datasets. We therefore propose a simplified version of the objective, where we estimate the second term of (10) with N sets of random samples from all modalities. Specifically, we can precompute the following $M \times N$ random index matrix J:

$$J = \begin{bmatrix} j_{11} & j_{12} & \cdots & j_{1N} \\ j_{21} & j_{22} & \cdots & j_{2N} \\ \vdots & & & \vdots \\ j_{M1} & j_{M2} & \cdots & j_{MN} \end{bmatrix}, \tag{12}$$

where each entry of J is a random integer taken from range $[1, N_m]$. We can then replace the second term of (10) random samples selected by the indices in J, giving us

$$\mathcal{L}_C(\boldsymbol{x}_{1:M}^{(i)}) \approx -\log p_\Theta(\boldsymbol{x}_{1:M}^{(i)}) + \left(\log \sum_{n=1}^{N} p_\Theta(\boldsymbol{x}_1^{(J_{1n})}, \boldsymbol{x}_2^{(J_{2n})}, \cdots \boldsymbol{x}_M^{(J_{Mn})}) \right). \tag{13}$$

The number of times ELBO needs to be computed is now $\mathcal{O}(N)$, and is no longer relevant to the number of modalities M.

We can now also generalise the final objective in (4) to M modalities:

$$\mathcal{L}(\boldsymbol{x}_{1:M}^{(i)}) = -\gamma \text{ELBO}(\boldsymbol{x}_{1:M}^{(i)}) + \left(\underset{n \in N}{\texttt{logsumexp}} \, \text{ELBO}(\boldsymbol{x}_1^{(J_{1n})}, \boldsymbol{x}_2^{(J_{2n})}, \cdots \boldsymbol{x}_M^{(J_{Mn})}) \right). \tag{14}$$

D THE INEFFECTIVENESS OF TRAINING WITH $\mathcal{L}_C$ ONLY

We demonstrate why training with the contrastive loss proposed in (3) is ineffective, and why additional ELBO term is needed for the final objective. As we show in Figure 9, when training with $\mathcal{L}_C$ only, while the contrastive loss (green) quickly drops to zero, both term ① and ② in (3) also reduces drastically. This means the joint marginal likelihood of any generation $\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ is small regardless the relatedness of $(\boldsymbol{x}, \boldsymbol{y})$.

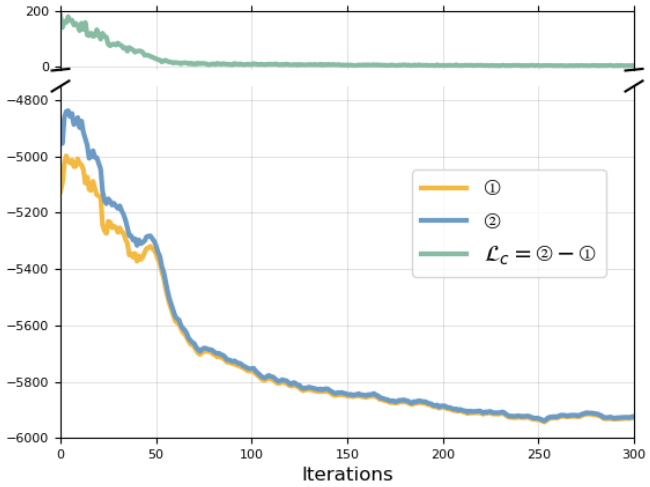


Figure 9: First 300 iterations of training using contrastive loss $\mathcal{L}_C$ only.

In comparison, we also plot the training curve for model trained on the final objective in (4), which upweights term ① in (3) by γ. We see in Figure 10 that by setting $\gamma = 2$, the joint marginal likelihood (yellow and blue curve) improves during training, while $\mathcal{L}_C$ (green curve) gets minimised.

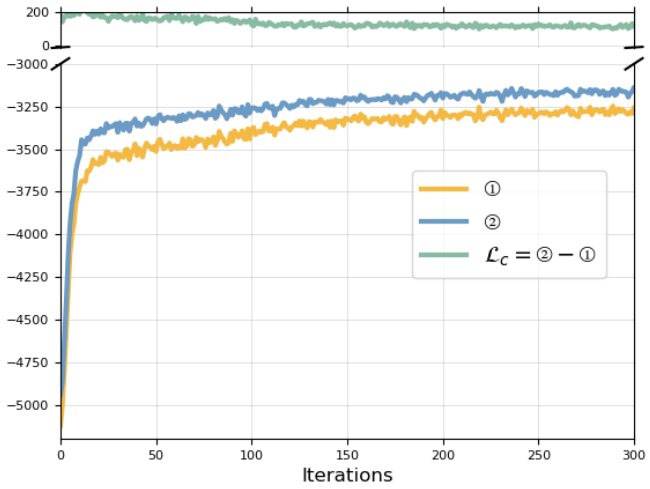


Figure 10: First 300 iterations of training with final loss $\mathcal{L}$, where $\gamma = 2$.

E ABLATION STUDY OF γ

In § 3, we specified that γ needs to be greater than 1 to offset the negative effect of minimising ELBO through term ② in (4). Here, we study the effect of γ in details.

Figure 11 compares latent accuracy, cross coherence and joint coherence of MMVAE on MNIST-SVHN dataset trained on different values of γ. Note that here we only consider cases where $\gamma \geq 1$, since the minimum value of γ is 1. In this case, the loss reduces to the original contrastive objective in (3).

A few interesting observations from the plot are as follows: First, when $\gamma = 1$, the model is trained using the contrastive loss only, and as we showed is an ineffective objective for generative model learning. This is verified again in Figure 11 — when $\gamma = 1$, both coherences and latent accuracies take on extremely low values; interestingly, there is a significant boost of performance across all metrics by simply increasing the value of γ from 1 to 1.1; after that, as the value of γ increases, performance on most metrics decreases monotonically (joint coherence being the only exception), and eventually converges to baseline MMVAE (dotted lines in Figure 11). This is unsurprising, since the final objective in (4) reduces to the original joint ELBO as γ approaches infinity.

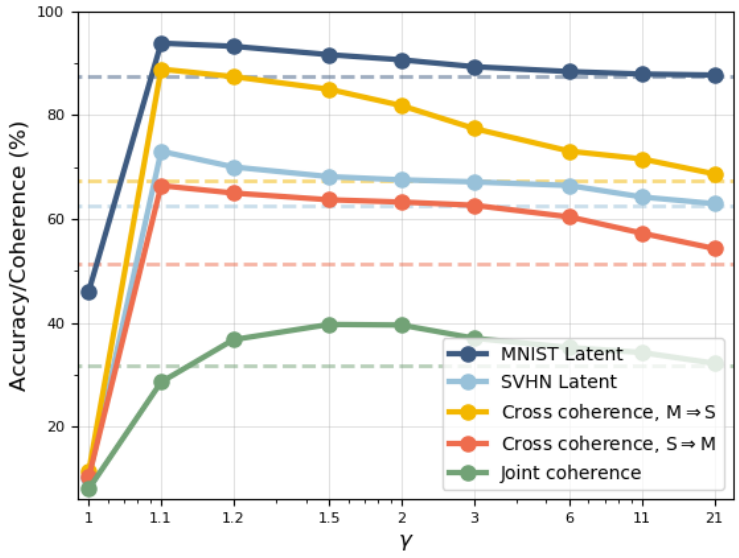


Figure 11: Performance on different metrics for different values of γ. Dotted lines represents the performance of baseline MMVAE.

Figure 11 seems to suggest that 1.1 is the optimal value for hyperparameter γ, however close inspection of the qualitative generative results shows that this might not be the case. See Figure 12 for a comparison of the model's generation between MMVAE models trained on (from left to right) $\gamma = 1.1$, $\gamma = 2$ and $\gamma = +\infty$ (i.e. original MMVAE). Although $\gamma = 1.1$ yields model with high coherence scores, we can clearly see from the left-most column of Figure 12 that the generation of the model seems deprecated, especially for the SVHN modality, where the backgrounds of model's generation appear to be unnaturally spotty and deformed. This problem is mitigated by increasing γ — as shown in Figure 11, the image generation quality of $\gamma = 2$ (middle column) is not visibly different from that of $\gamma = +\infty$ (right column).

To verify this observation, we also compute the marginal log likelihood $\log p_\Theta(\boldsymbol{x}, \boldsymbol{y})$ to quantify the quality of generations. We compute this for all γs considered in Figure 11, and take the average over the entire test set. From the results in Figure 13, we can see a significant increase of the log likelihood between $\gamma = 1.1$ to $\gamma = 1$. This gain in image generation quality then slows down as γ further increases, and as all other metrics converges to the original MMVAE model.

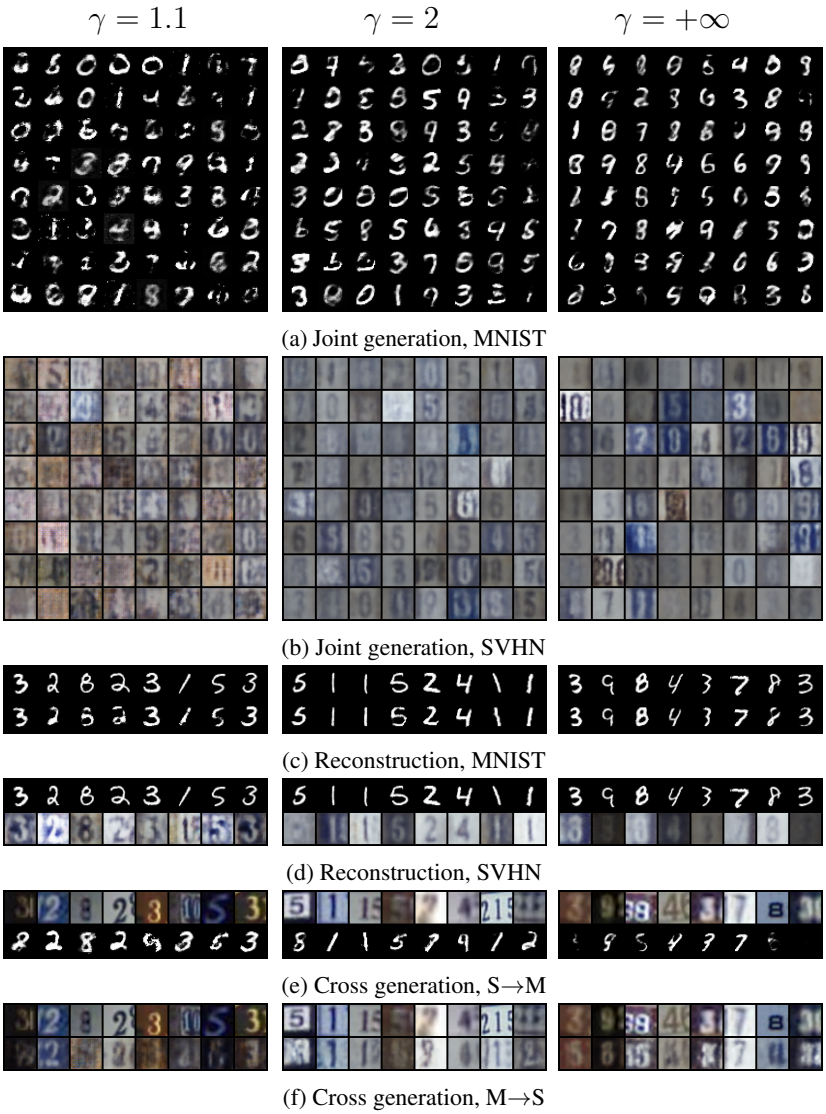


Figure 12: Generations of MMVAE model trained using the final contrastive objective, with (from left to right) $\gamma = 1.1$, 2 and $+\infty$. Note in (c), (d), (e), (f), the top rows are the inputs and the bottom rows are their corresponding reconstruction/cross generation.

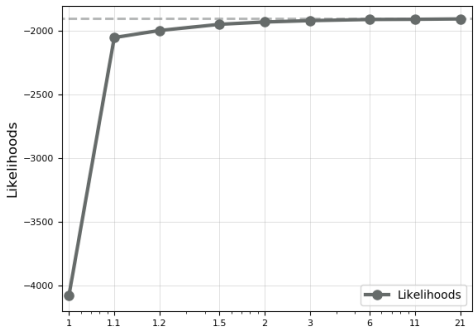


Figure 13: Performance on different metrics for different values of γ. Dotted lines represents the performance of baseline MMVAE.

F ARCHITECTURE

We use architectures listed in Table 3 for the unimodal encoder and decoder for MMVAE, MVAE and JMVAE. For JMVAE we use an extra joint encoder, the architecture of which is described in Table 4.

Encoder	**Decoder**
Input $\in \mathbb{R}^{1x28x28}$	Input $\in \mathbb{R}^{L}$
FC. 400 ReLU	FC. 400 ReLU
FC. L, FC. L	FC. 1 x 28 x 28 Sigmoid

(a) MNIST dataset

Encoder
Input $\in \mathbb{R}^{3x32x32}$
4x4 conv. 32 stride 2 pad 1 & ReLU
4x4 conv. 64 stride 2 pad 1 & ReLU
4x4 conv. 128 stride 2 pad 1 & ReLU
4x4 conv. L stride 1 pad 0, 4x4 conv. L stride 1 pad 0

Decoder
Input $\in \mathbb{R}^{L}$
4x4 upconv. 128 stride 1 pad 0 & ReLU
4x4 upconv. 64 stride 2 pad 1 & ReLU
4x4 upconv. 32 stride 2 pad 1 & ReLU
4x4 upconv. 3 stride 2 pad 1 & Sigmoid

(b) SVHN dataset.

Encoder	**Decoder**
Input $\in \mathbb{R}^{2048}$	Input $\in \mathbb{R}^{L}$
FC. 1024 ELU	FC. 256 ELU
FC. 512 ELU	FC. 512 ELU
FC. 256 ELU	FC. 1024 ELU
FC. L, FC. L	FC. 2048

(c) CUB image dataset.

Encoder
Input $\in \mathbb{R}^{1590}$
Word Emb. 128
4x4 conv. 32 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 conv. 64 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 conv. 128 stride 2 pad 1 & BatchNorm2d & ReLU
1x4 conv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
1x4 conv. 512 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
4x4 conv. L stride 1 pad 0, 4x4 conv. L stride 1 pad 0

Decoder
Input $\in \mathbb{R}^{L}$
4x4 upconv. 512 stride 1 pad 0 & ReLU
1x4 upconv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
1x4 upconv. 128 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
4x4 upconv. 64 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 upconv. 32 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 upconv. 1 stride 2 pad 1 & ReLU
Word Emb.^{T} 1590

(d) CUB-Language dataset.

Table 3: Unimodal encoder and decoder architectures.

Encoder
Input $\in \mathbb{R}^{3x32x64}$
4x4 conv. 32 stride 2 pad 1 & ReLU
4x4 conv. 64 stride 2 pad 1 & ReLU
4x4 conv. 128 stride 2 pad 1 & ReLU
1x4 conv. 128 stride 1x2 pad 0x1 & ReLU
4x4 conv. L stride 1 pad 0, 4x4 conv. L stride 1 pad 0

(a) MNIST-SVHN dataset.

Encoder
Input $\in \mathbb{R}^{1x32x192}$
4x4 conv. 32 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 conv. 64 stride 2 pad 1 & BatchNorm2d & ReLU
4x4 conv. 128 stride 2 pad 1 & BatchNorm2d & ReLU
1x4 conv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
1x4 conv. 512 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
4x4 conv. L stride 1 pad 0, 4x6 conv. L stride 1 pad 0

(b) CUB Image-Caption dataset.

Table 4: Joint encoder architectures.