Efficient Modality Translation via Arbitrary Conditioning and Wasserstein Regularization

Tomas Tokar[1, 2], **Scott Sanner**[2, 3]

[1]Wondeur AI
[2]University of Toronto
[3]Vector Institute for AI

Abstract

The central challenge in multimodal generative modeling lies in accurately approximating the joint data distribution, even when some modalities are missing. Existing multimodal VAEs solve this by designing increasingly complex encoding architectures, relying on modality-specific encoders, factorized posteriors, and custom inference procedures. This restricts their ability to capture relations among modalities by amortizing the encoding parameters. We challenge this paradigm by introducing a model trained for arbitrary conditioning, i.e., generating any modality given a subset of observed modalities and a logical index indicating which modalities are present or missing. This enables a single unified encoder to handle any subset of modalities while capturing inter-modal relationships via a compact, shared posterior. We find that to work efficiently in the multimodal setup, arbitrary conditioning requires replacing the KL divergence with Wasserstein regularization, which allows more dispersed latent embeddings to support learning over diverse data and modality subsets. This key insight exposes a critical deficiency in existing methods, which rely on KL regularization that tends to concentrate individual embeddings near the standard Gaussian prior, despite coming from very diverse subsets of multimodal inputs. We prove that Wasserstein regularization ensures that the aggregate latent distribution – spanning all conditioning subsets – aligns with the prior without requiring mixture models or auxiliary inference tricks. Empirically, the proposed model improves cross-modal generation and yields better reconstructions than state-of-the-art multimodal VAEs.

Introduction

The goal of multimodal data modeling is to learn high-quality representations (Bengio, Courville, and Vincent 2013) of multimodal data to facilitate downstream supervised tasks (Guo, Wang, and Wang 2019), while simultaneously to accurately approximate the data distribution to support generative tasks (Suzuki and Matsuo 2022).

To meet this dual goal, considerable attention has been paid to multimodal adaptations of variational autoencoders (VAEs) (Kingma and Welling 2014). Current multimodal VAEs use architectural designs that allow them to encode any subset of modalities into a shared posterior distribution, and thus handle missing modalities during inference and enable cross-modal generation, often referred to as *modality translation* (Suzuki and Matsuo 2022; Żelaszczyk and Mańdziuk 2023).

The encoding architectures used in these models involve complex inference via unimodal encoders, restricting the ability of the neural networks to capture relations among modalities by amortizing the encoding parameters. Moreover, these models use Kullback-Leibler (*KL*) regularization encouraging the inputs to be encoded to posterior distributions that approximate the selected prior (usually a standard Gaussian distribution), inadvertently forcing the latent embeddings to concentrate close to each other, despite coming from diverse inputs, or subsets of observed modalities – a critical limitation that remains largely overlooked.

To address these deficiencies, we propose to treat cross-modal generation as *arbitrary conditioning* (Ivanov, Figurnov, and Vetrov 2018; Belghazi, Oquab, and Lopez-Paz 2019), where the embeddings are conditioned on the logical index indicating which modalities are observed. Furthermore, we propose to replace KL-divergence with Wasserstein regularization (Arjovsky, Chintala, and Bottou 2017; Tolstikhin et al. 2018), which permits the latent codes to be far from each other if coming from distinct inputs or different subsets of observed modalities. Together, but not separately, these adaptations, allow us to flexibly handle any combination of observed modalities and learn their joint representation in a principled manner.

We prove that Wasserstein regularization imposes control over the overall latent distribution, spanning across data inputs and, critically, also across the combinations of observed modalities, without the need for non-standard priors (e.g., Gaussian mixtures) or complex training procedures. This is a key insight that extends the original argument for the use of Wasserstein autoencoders (WAEs) (Rubenstein, Schoelkopf, and Tolstikhin 2018; Tolstikhin et al. 2018) and highlights WAEs as particularly well suited for arbitrary conditioning.

The proposed method is denoted as Multimodal Arbitrary Conditioning with Wasserstein Autoencoder (MAC-WAE) (cf. Figure 1)[1]. To demonstrate the versatility and robustness of MAC-WAE across diverse data, we conducted exper-

[1]The project code and supplementary materials are publicly available at: https://github.com/tomastokar/MACWAE

iments on four multimodal datasets, including image, text, and audio modalities. Compared to state-of-the-art baselines, MAC-WAE achieves significantly better cross-modal translation and yields learned representations that lead to more faithful input reconstructions.

Preliminaries

Let $\mathbf{x}_i$ denote the i-th data instance, comprising M inputs from different modalities: $\mathbf{x}_i = \{\mathbf{x}_i^{(m)}\}_{m=1}^M$. Assume that each input $\mathbf{x}^{(m)}$ comes from a distinct modality-specific input space $\mathcal{X}^{(m)}$: $\mathbf{x}^{(m)} \in \mathcal{X}^{(m)}$. Furthermore, let $\mathbf{x}_i^{(\mathcal{A})} = \{\mathbf{x}_i^{(m)}\}_{m\in\mathcal{A}}$ denote an instance comprising only a subset of modalities $\mathcal{A}$, so that: $\mathcal{A} \subset \mathcal{P}(M)$, where $\mathcal{P}(M)$ denotes the power set of modalities; and let $\mathbf{x}_i^{(\setminus m)} = \{\mathbf{x}_i^{(j)}\}_{j\neq m}^M$ be a specific case of an input comprising all modalities except the m-th modality. Finally, consider $\mathbf{x} \sim p(\mathbf{x})$, and $\mathbf{x}^{(m)} \sim p(\mathbf{x}^{(m)})$, where $p(\mathbf{x})$ denotes the joint data distribution, and $p(\mathbf{x}^{(m)})$ denotes the marginal distribution of a given modality m; and let $p(\mathbf{x}^{(\mathcal{Q})}|\mathbf{x}^{(\mathcal{V})})$ denote the conditional probability of $\mathbf{x}^{(\mathcal{Q})}$ given the inputs $\mathbf{x}^{(\mathcal{V})}$, assuming $\mathcal{Q} \cap \mathcal{V} = \emptyset$.

Related Work

Multimodal VAEs

Multimodal VAEs are an extension of traditional VAEs (Kingma and Welling 2014). In most cases, multimodal VAEs learn to model the multimodal data distribution conditioned on a shared latent variable $\mathbf{z}$: $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ and $\mathbf{z} \sim p(\mathbf{z})$, where θ represents the model parameters and the $p(\mathbf{z})$ is a prior distribution, which is usually assumed to be a normal distribution: $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

It is usually assumed that the likelihood factorizes across modalities, so that $p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{m=1}^M p_{\theta_m}(\mathbf{x}^{(m)}|\mathbf{z})$, where p_{θ_m} is selected based on the statistical properties of $\mathbf{x}^{(m)}$. The assumed generative process can be then described as:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \prod_{m=1}^{M} p_{\theta_m}(\mathbf{x}^{(m)}|\mathbf{z}) \tag{1}$$

The multimodal VAEs are trained to maximize the lower bound of the marginal distribution $p(\mathbf{x})$:

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\sum_{m=1}^{M} \log p_\theta(\mathbf{x}^{(m)}|\mathbf{z})\Big] - D_{KL}(q_\phi||p(\mathbf{z})) \tag{2}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is an approximation of the true posterior distribution $p(\mathbf{z}|\mathbf{x})$, and ϕ denotes the associated model parameters.

The most important challenge with multimodal VAEs is how to handle missing modalities, which is a prerequisite for being able to perform modality translation. There are three main approaches to this, which we survey as follows.

Surrogate unimodal inference models To handle missing modalities, (Suzuki, Nakayama, and Matsuo 2017) introduced *surrogate unimodal inference distributions* $q_{\phi_m}(\mathbf{z}|\mathbf{x}^{(m)})$, alongside the joint posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, while extending the objective in Eq. 2 for minimizing the KL divergence between the surrogates and the joint posterior. This way, each unimodal distribution was encouraged to approximate the joint distribution and hence the model was optimized to encourage conditional generation. Multiple modifications of this approach were later proposed (Vedantam et al. 2017; Korthals et al. 2019; Senellart, Chadebec, and Allassonnière 2023).

Mixture-based posterior models Mixture-based multimodal VAEs introduced a modified multimodal variational posterior $q_\phi(\mathbf{z}|\,\mathbf{x})$, the functional form of which can be generalized as follows (Javaloy, Meghdadi, and Valera 2022):

$$q_\phi(\mathbf{z}|\mathbf{x}) = \frac{1}{|A|} \sum_{A\in\mathcal{A}} q_A(\mathbf{z}|\mathbf{x}^{(\mathbf{A})}) \tag{3}$$

where $\mathcal{A}$ is a subset of $\mathcal{P}(M)$: $\mathcal{A} \subseteq \mathcal{P}(M)$, and $q_A(\mathbf{z}|\mathbf{x}^{(\mathbf{A})})$ is product of unimodal posteriors $q_{\phi_m}(\mathbf{x}|\mathbf{x}^{(m)})$, sometimes called "experts", which are parameterized by neural networks and usually selected to have Gaussian form:

$$q_A(\mathbf{z}|\mathbf{x}^{(A)}) \propto \prod_{m\in A} q_{\phi_m}(\mathbf{x}|\mathbf{x}^{(m)}) \tag{4}$$

The individual methods (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2021) are then a special cases of the above generalization, distinguished by their choice of $\mathcal{A}$: **MVAE** (Wu and Goodman 2018) selects $\mathcal{A} = \{\{1, \dots M\}\}$ so that the eq. 3 converts to $q_\phi(\mathbf{z}|\mathbf{x}) \propto \prod_{m=1}^M q_{\phi_m}(\mathbf{x}|\mathbf{x}^{(m)})$, i.e. "product of experts" (Hinton 2002). **MMVAE** (Shi et al. 2019) selects $\mathcal{A} = \{\{1\}, \dots \{M\}\}$ so that eq. 3 reduces to $q_\phi(\mathbf{z}|\mathbf{x}) \propto \frac{1}{M}\sum_{m=1}^M q_{\phi_m}$, i.e. "mixture of experts". **MoPoE** (Sutter, Daunhawer, and Vogt 2021) is the exhaustive case, where $\mathcal{A} = \mathcal{P}(M)$, the posterior in eq. 3 becomes "mixture of products of experts", bridging the previous two approaches.

Hierarchical models Hierarchical models introduce additional latent variables, which are organized in the hierarchal structure, so that hierarchically lower-level variables are conditioned on the upper-level ones. In the simplest case, one assumes modality-specific latent variables $\mathbf{z}^{(m)}$ encode individual modalities, each conditioned on the joint latent variable $\mathbf{z}$, so that the generative process extends to a three-step process:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{z}^{(1:M)}) = p(\mathbf{z}) \prod_{m=1}^{M} p(\mathbf{z}^{(m)}|\mathbf{z}) p(\mathbf{x}^{(m)}|\mathbf{z}^{(m)}) \tag{5}$$

The modality-specific encoders and decoders can be learned either through independent pre-training or together with the joint encoder and decoder.

In the presence of missing modalities, hierarchical models either fill the missing modality-specific latent variables by sampling from their prior (Ma et al. 2020; Nazabal et al.

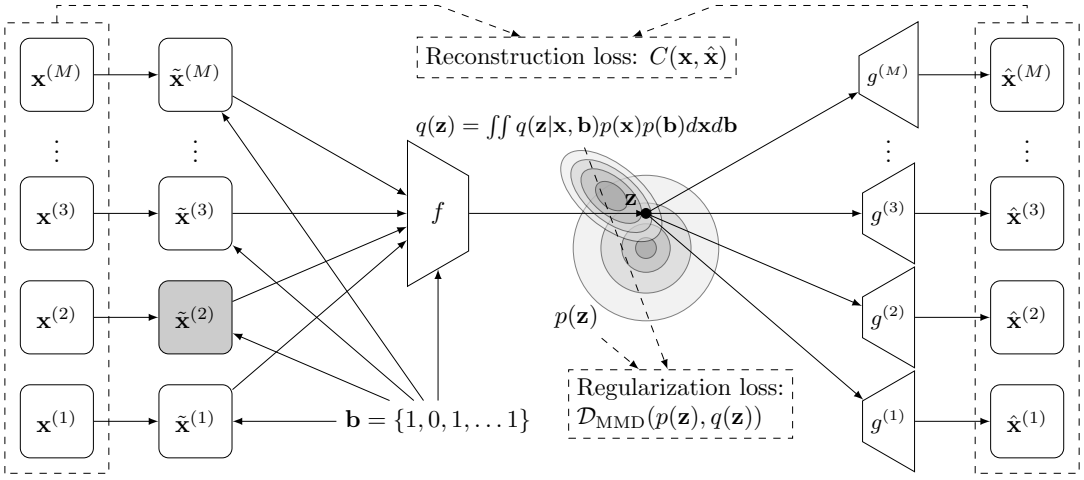


Figure 1: MAC-WAE: A *conditioning index* $\mathbf{b}$ is sampled from $p_{\mathbf{b}}$ to indicate which modalities are observed and which are missing. The inputs $\mathbf{x}^{(m)}$ and b are jointly encoded via encoder f to produce latent code $\mathbf{z}$, which is then decoded by the modality-specific decoders $g^{(m)}$. The model aims to simultaneously minimize the reconstruction loss C and the regularization loss $\mathcal{D}_{\text{MMD}}$ between the prior $p(\mathbf{z})$ and the continuous mixture $q(\mathbf{z})$.

2020; Vasco, Melo, and Paiva 2020), or they rely on joint encoder architectures that are robust against the missing inputs; such as GraphSAGE (Hamilton, Ying, and Leskovec 2017) employed by **Nexus** (Vasco et al. 2022).

Wasserstein Autoencoders

Wasserstein autoencoders (WAEs) (Tolstikhin et al. 2018) are a type of generative model that combine autoencoders with the principles of optimal transport theory, particularly the Wasserstein distance. Instead of using the D_{KL} for regularization as in VAEs, they minimize the Wasserstein distance (Rüschendorf 1985) between the encoded latent distributions and a prior distribution, thus encouraging the continuous mixture of posteriors: $q(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})d\mathbf{x}$ to match the selected prior. This provides better control over the generative process and often leads to improved generative quality and training stability (Rubenstein, Schoelkopf, and Tolstikhin 2018).

In the original work (Tolstikhin et al. 2018), the authors proposed the penalty based on maximum mean discrepancy (MMD) (Smola, Gretton, and Borgwardt 2006), which measures the distance between the prior $p(\mathbf{z})$ and $q(\mathbf{z})$ using distribution-embedding kernels:

$$\mathcal{D}_{\text{MMD}}(p(\mathbf{z}), q(\mathbf{z})) = \mathbb{E}_{\mathbf{z},\widetilde{\mathbf{z}}\sim p(\mathbf{z})}[k(\mathbf{z}, \widetilde{\mathbf{z}})] + \mathbb{E}_{\mathbf{z},\widetilde{\mathbf{z}}\sim q(\mathbf{z})}[k(\mathbf{z}, \widetilde{\mathbf{z}})] - 2\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}),\widetilde{\mathbf{z}}\sim q(\mathbf{z})}[k(\mathbf{z}, \widetilde{\mathbf{z}})] \quad (6)$$

where k is a positive-definite kernel. While MMD is known to perform well on high-dimensional standard normal distributions (Gretton et al. 2012), there are several alternative penalties that can be used for this purpose (cf. Section).

Surprisingly, the application of WAEs within a multimodal setup remains largely unexplored. In (Tsai et al. 2017), the authors proposed factorized multimodal latent representations, consisting of private (modality-specific) and shared variables, which were regularized by Wasserstein distance. However, the model requires designated neural encoders, each to handle a single missing modality, rendering this approach unsuitable for general modality transfer.

Somewhat related are the works of (Tian and Engel 2019) and (Schonfeld et al. 2019), where the authors used Wasserstein distance to match posterior distributions of two classical VAEs. Conversely, (Mahajan et al. 2019) pre-trained a WAE to obtain Gaussian-regularized image and text representations, which were then aligned using a contrastive learning approach. However, these works are limited to bimodal cases and none of them allow for joint modeling of the true multimodal data distribution.

Arbitrary Conditioning

Generative models, such as VAEs, aim to model the data distribution $p(\mathbf{x})$ and so, hypothetically, these models should also be able to approximate any arbitrary conditional distribution $p(\mathbf{x}_u|\mathbf{x}_o)$, where $\mathbf{x}_u$ and $\mathbf{x}_o$ indicate unobserved and observed portions of the data, respectively. However, in practice, deriving the conditional by marginalizing the joint model distribution is often not feasible, since the analytically solutions are generally unavailable and computational approximations are often intractable. For that reason, we may seek to model such arbitrary conditional distributions directly – a task referred to as *arbitrary conditioning*.

Various types of generative methods were previously modified for arbitrary conditioning. These modifications include GANs (Belghazi, Oquab, and Lopez-Paz 2019), normalizing flows (Li, Akbar, and Oliva 2020), autoregressive models (Strauss and Oliva 2021) as well as VAEs (Ivanov, Figurnov, and Vetrov 2018; Ma et al. 2019; Strauss and Oliva 2022). However, *none of these models were intended for multimodal data; nor can they be easily adapted for multimodal applications*. This motivated development of the pro-

posed method.

Multimodal Arbitrary Conditioning with Wasserstein Autoencoder

Problem Statement

VAE objective seeks to maximize the similarity (by minimizing the KL-divergence in Eq. 2) between the prior and each individual posterior $q(\mathbf{z}|\mathbf{x}_i, \mathbf{b}_j)$, thus forcing the latent codes to remain close to each other despite coming from different instances. This deficiency of VAEs is well known in various contexts (Makhzani et al. 2015; Hoffman and Johnson 2016; Alemi et al. 2018; Tomczak and Welling 2018).

This property of VAEs becomes especially problematic when used to model arbitrary conditional distributions; as the latent codes coming from distinct subsets of the observed inputs are compressed together. Even more critical this becomes in a multimodal setup, where good approximation of the conditional distributions requires subsets of observed modalities to be mapped into vastly diverse latent codes. We thus argue that multimodal VAEs suffer by inherent deficiency when modeling arbitrary cross modal conditionals.

Rationale

Inference To encourage the model to learn to infer latent representations of the inputs in spite of missing modalities, we propose the following inference schema (cf. Figure 1). The binary index $\mathbf{b} \in \{0, 1\}^M$ is sampled from the distribution $p_{\mathbf{b}}$, which serves as a *conditioning index*. By default, the distribution $p_{\mathbf{b}}$ is selected as follows:

$$p(\mathbf{b}) = \begin{cases} 0, & \text{if } \mathbf{b} = \mathbf{0} \\ \dfrac{1}{1 - p_0^M} \displaystyle\prod_{i=1}^{M} p_0^{1-b_i}(1 - p_0)^{b_i}, & \text{otherwise} \end{cases} \tag{7}$$

i.e., a joint Bernoulli distribution that is constrained to avoid all values of b to be zeros that guarantees that at least one modality is always observed; where p_0 is a hyper-parameter of the model.

The modality m is considered observed in the given batch if $b_m = 1$, and missing otherwise; in which case the input $\mathbf{x}^{(m)}$ is set to zero (cf. Section). The inputs are then passed to the joint neural encoder f_ϕ that implements the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{b})$.

Generative process The generative process is identical to the one described in Eq. 1. The code $\mathbf{z}$ is sampled from $q_\theta(\mathbf{z}|\mathbf{x}, \mathbf{b})$ and subsequently passed to decoder g, which consists of the set of M independent modality-specific neural decoders $g^{(m)}$ with parameters θ_m: $g_\theta = \{g^{(1)}_{\theta_1}, \dots, g^{(M)}_{\theta_M}\}$. The reconstruction $\hat{\mathbf{x}} = g_\theta(\mathbf{z})$ is then a set of the modality-specific reconstructions: $\hat{\mathbf{x}} = \{g^{(m)}_{\theta_m}(\mathbf{z})\}_{m=1}^M$.

Loss computation The aim is to minimize the reconstruction loss measured by the multimodal *cost function* C: $C(\mathbf{x}, g_\theta(\mathbf{z})) = \sum_{m=1}^M c(\mathbf{x}, g^{(m)}_{\theta_m}(\mathbf{z}))$, where c is the L_2 norm. Simultaneously, we aim to minimize the regularization penalty $\mathcal{D}$, the choice of which and its subsequent theoretical implications are elaborated in the following sections.

For each batch, the encoding and decoding steps are repeated K times, each time using different conditioning index $\mathbf{b}$. The incurred reconstruction and regularization losses are summed into a reconstruction and regularization batch losses respectively, which are then aggregated into a total batch loss, where hyper-parameter β controls the strength of regularization (Higgins et al. 2017; Burgess et al. 2018) (cf. Algorithm 1).

Regularization Metric

Similarly to (Tolstikhin et al. 2018), we adopted the penalty based on maximum mean discrepancy (MMD) (Smola, Gretton, and Borgwardt 2006), using radial basis function (RBF) kernel (Benoudjit and Verleysen 2003). We estimate $\mathcal{D}_{\text{MMD}}$ between the standard Gaussian prior $p(\mathbf{z})$ and marginal posterior $q(\mathbf{z}|\mathbf{b})$, which is essentially a continuous mixture of posteriors across the instances: $q(\mathbf{z}|\mathbf{b}) = \int_x q(\mathbf{z}|\mathbf{x}, \mathbf{b})p(\mathbf{x})d\mathbf{x}$ (Tolstikhin et al. 2018). To estimate $\mathcal{D}_{\text{MMD}}$ we use the U-statistic (Ferguson, Shapley, and MacQueen 2005), so the resulting regularization penalty is defined as

$$\mathcal{D}_{\text{MMD}}(p(\mathbf{z}), q_\phi(\mathbf{z}|\mathbf{b})) = \frac{1}{n-1}\sum_{l\neq j} k(\mathbf{z}_l, \mathbf{z}_j) + \frac{1}{n(n-1)}\sum_{l\neq j} k(\widetilde{\mathbf{z}}_l, \widetilde{\mathbf{z}}_j) - \frac{2}{n^2}\sum_{l,j} k(\mathbf{z}_l, \widetilde{\mathbf{z}}_j) \tag{8}$$

where $(\mathbf{z}_i)_{i=1}^n$ are samples drawn from the prior $p(\mathbf{z})$; $(\widetilde{\mathbf{z}}_i)_{i=1}^n$ are the latent codes of the inputs $\mathbf{x}_i$ for $i = 1, \dots n$ in the batch under the sampled index $\mathbf{b}$; i. e., essentially the samples drawn from $q_\phi(\mathbf{z}|\mathbf{b})$.

Theoretical Considerations

The learning objective of MAC-WAE is thus to minimize the following loss:

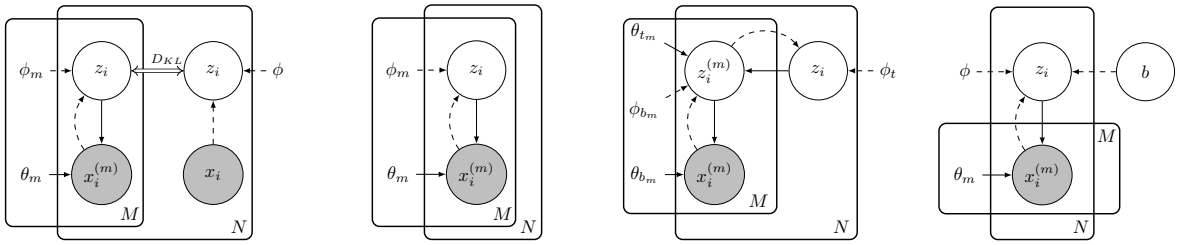
$$\mathcal{L}_{\text{MACWAE}} = \mathbb{E}_{p(\mathbf{b})}\Big[\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{b})}[C(\mathbf{x}, g(\mathbf{z}))] + \beta \cdot \mathcal{D}_{\text{MMD}}(p(\mathbf{z}), q(\mathbf{z}|\mathbf{b}))\Big] \tag{9}$$

Theorem 0.1. *For any fixed prior probability density function $p(\mathbf{z})$, the loss $\mathcal{L}_{MACWAE}$ incurred by the model is a subject to the following lower bound:*

$$\mathcal{L}_{MACWAE} \geq \mathbb{E}_{p(\mathbf{b})}\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{b})}[c(\mathbf{x}, g(\mathbf{z}))] + \beta \cdot \mathcal{D}_{MMD}(p(\mathbf{z}), q(\mathbf{z})) \tag{10}$$

where $q(\mathbf{z}) = \mathbb{E}_{p_{\mathbf{b}}}[q(\mathbf{z}|\mathbf{b})]$.

Importantly, from the Theorem 0.1 (proof is provided in Supplementary Materials) follows that minimizing the loss $\mathcal{L}_{\text{MACWAE}}$ leads to minimizing the sum of reconstruction loss incurred across the data and the conditioning indices (first term) and the regularization loss (second term). The minimization of the latter means to maximize the similarity between the prior $p(\mathbf{z})$ and the continuous mixture of posteriors: $q(\mathbf{z}) = \int_x \int_b q(\mathbf{z}|\mathbf{x}, \mathbf{b})p(\mathbf{x})p(\mathbf{b})d\mathbf{x}d\mathbf{b}$, going across the data and the conditioning indices.



(a) JMVAE (b) MoPoE/MVTCAE (c) Nexus (d) **MAC-WAE**

Figure 2: Graphical models of the four baselines: JMVAE (Suzuki, Nakayama, and Matsuo 2017), MoPoE (Sutter, Daunhawer, and Vogt 2021), MVTCAE (Hwang et al. 2021), Nexus (Vasco et al. 2022) and MAC-WAE. Shaded nodes indicate observed variables, white nodes indicate latent ones. Frames indicate the modalities plate (M) and instances plate (N). Solid and dashed arrows indicate generative and inference processes, respectively; and $\Leftrightarrow$ indicates Kullback-Leibler divergence between the distributions of the given variables. Note, only MAC-WAE exerts the inference jointly via a single multimodal encoder (z outside of M plate); allowing it to better capture cross-modal relations by amortizing the encoder parameters.

Theoretical benefits The Wasserstein regularization lends our method an important theoretical advantage with respect to the cross-modal generation, since *the resulting latent codes are allowed to be far away from each other, if coming from different instances and, more importantly, different conditioning indices*. This advantage is particularly beneficial in a multimodal setup, where different subsets of observed modalities are likely to result in vastly diverse latent codes. Furthermore, the model comes with only a single multimodal encoder that exerts the inference from all the observed modalities jointly. In contrast, most existing methods rely on unimodal encoders, losing the ability of the neural networks to capture relations among modalities by amortizing the parameters (cf. Figure 2).

Encoder and Decoders

An important, yet often underappreciated, property of WAEs is their permissibility for deterministic encoders, i.e., encoders that map the given input $\mathbf{x}_i$ to the exact same latent code $\mathbf{z}_i$, without the model losing its generative nature. The posterior distribution then has the form of a Dirac δ distribution, instead of a traditional Gaussian. Unless stated otherwise, we use deterministic encoders in this work.

For the sake of epistemological clarity of our research, we sought as parsimonious of an encoding and decoding architecture as possible (cf. Supplementary Materials). Outputs of the modality specific encoders were concatenated and passed to a simple joint encoder; consisting of a single hidden layer, ReLU-activated, multi-layer perceptron (MLP) (Noriega 2005) to produce the joint posterior. To accommodate missing modalities, inputs from missing modalities were replaced with zeros.

Experimental Design

Research questions Our experiments address the following research questions: [**RQ1**] Does the MAC-WAE outperform state-of-the-art (SOTA) baselines with respect to the three *tasks*: (a) reconstruction quality; (b) *many-to-one* modality translation, i.e. the ability to estimate the single missing modality given the remaining ones collectively; (c) *one-to-one* modality translation, i.e. the ability to estimate a single modality from other modalities individually; (d) generative quality. [**RQ2**] Does the use of the Wasserstein regularization by MAC-WAE significantly improve the model performance compared to the use of *KL* regularization?

Baselines To provide meaningful and diverse baselines we selected the 4 methods, which constitute canonical representatives of their respective research avenues (cf. Figure 2): **JMVAE** (Suzuki, Nakayama, and Matsuo 2017) representing multimodal VAEs with unimodal surrogate inference. **MoPoE** (Sutter, Daunhawer, and Vogt 2021) representing an important generalization of the mixture-based posterior VAEs. **NEXUS** (Vasco et al. 2022) representing hierarchical multimodal VAEs. **MVTCAE** (Hwang et al. 2021) is an information-theoretic model that adopts a product-of-experts joint encoder, but uses the variational information bottleneck as an additional form of regularization.

We acknowledge the existence of several other works on multimodal VAEs (e.g., (Palumbo, Daunhawer, and Vogt 2023; Sutter et al. 2024)). However, to the best of our knowledge, these approaches build upon existing baseline architectures by introducing additional enhancements that are equally applicable to our proposed method. Therefore, we focus our comparisons on the selected canonical baselines and do not include these variants in our evaluation.

Ablated models To evaluate the benefits of Wasserstein regularization as compared to traditional KL-divergence we ablated the proposed model by replacing the $\mathcal{D}_{\text{MMD}}$ in the Equation 9 with KL-divergence while simultaneously replacing our deterministic encoder with the commonly used VAE stochastic encoder with Gaussian posterior. Hereafter, we will refer to this ablated model as **MAC-VAE**. To evaluate the contribution of arbitrary conditioning, we proposed a second ablated model, which uses Wasserstein regularization but lacks arbitrary conditioning, hence reducing the MAC-WAE model to a masked Wasserstein autoencoder, which we refer to as **M-WAE**.

Evaluation metric The quality of the reconstruction and modality translation was measured by modality-specific evaluation metrics, namely: statistical accuracy (ACC) for the categorical modalities, root-mean-squared error (RMSE) for the numerical modalities, structural similarity index (SSIM) (Wang et al. 2004) for the images and BLEU-3 score (BLEU3) (Papineni et al. 2002) for text (cf. Supplementary Materials).

Datasets In our experiments we used only the following publicly available datasets: **PolyMNIST**, instances of which consist of a set of 5 MNIST images, where each image in a set represents the same digit, but appears with different backgrounds and in different handwriting styles, which introduce a multimodal aspect of the data (Sutter, Daunhawer, and Vogt 2021). **MHD** is a collection of multimodal handwritten digits, combining images, motion trajectory, sound and semantic information from digit handwriting (Vasco et al. 2022). **CUB-Captions** consists of images of birds paired with matching linguistic descriptions and the associated species labels (Netzer et al. 2011; Shi et al. 2019). **CelebA** is a large-scale face dataset containing celebrity images across diverse background, poses and lightning conditions, with each image annotated for 40 binary attributes, which we treated as individual modalities (Liu et al. 2015).

Results

The models were compared across three tasks and four datasets comprising 14 modalities in total. This resulted in 42 experimental *outcomes* (cf. Supplementary Table 4). The MAC-WAE outperformed the baselines and the ablated model in 29 of these outcomes, i.e., in approximately 70%.

Average performance ranking To obtain task-level performance measures, the obtained results were aggregated by computing the *average performance ranking*. The models were ranked based on their performance on the individual modalities, with the best-performing model assigned a value of 1 and the worst-performing model assigned a value of 0. The obtained rankings were then averaged for each task (cf. Figure 3). The MAC-WAE achieved the highest ranking across all three tasks, and so we conclude that MAC-WAE outperforms the given baselines (RQ1, *a–c*).

Win-loss matrix To provide compact pair-wise model comparison across the experimental outcomes, the obtained results were further conveyed as a win-loss matrix (cf. Figure 4). The MAC-WAE significantly ($p\text{–}value < 0.05$, binomial test) outperformed all the baselines as well as the ablated model. This demonstrates that the use of Wasserstein regularization by MAC-WAE significantly improves the model performance compared to the *KL* regularization (RQ2).

Qualitative analysis of the produced embeddings We visualized UMAP-processed (McInnes et al. 2018) latent representations conditioned on various subsets of modalities across the test portion of MHD dataset (multimodal handwritten digits), as produced by the MAC-WAE and by the non-Wasserstein ablation model MAC-VAE. The obtained results show qualitative differences between the latent representations produced under Wasserstein and KL regularization (Figure 5).

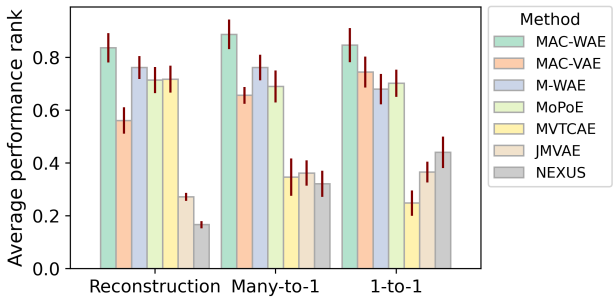


Figure 3: *Average performance rank* in reconstruction quality, *many-to-1* and *1-to-1* modality translation. Rankings were computed by assigning a score of 1 to the best-performing model and 0 to the worst-performing model across all datasets and modalities. The scores were then averaged to obtain the final ranking per task, providing a comparative evaluation of the models' performance.

Robustness to the choice of hyper-parameter β We investigated the sensitivity of model performance to the strength of the regularization term, controlled by the hyper-parameter β, for both MAC-WAE and baseline models across three representative tasks: reconstruction, many-to-one modality translation, and one-to-one modality translation. These evaluations were conducted on the PolyMNIST, CUB, and MHD datasets. The results demonstrate that MAC-WAE exhibits overall greater robustness to variations in β compared to the baseline models, maintaining more stable performance across range of regularization strengths (cf. Supplementary Figures 2–4).

Discussion

The central challenge in multimodal data modeling lies in learning high-quality representations of multimodal data while simultaneously enabling effective cross-modal generation, or modality translation. We proposed framing cross-modal generation as a problem of arbitrary conditioning and leveraging Wasserstein autoencoders (WAEs) as the underlying generative model. As we showed, WAEs are particularly well-suited to this task due to their unique regularization objective. By employing arbitrary conditioning within the WAE framework, our approach creates a unified model capable of excelling at both, multimodal representation learning and arbitrary cross-modal generation.

Notably, the closest contender to our model is MoPoE, which derives its learning capacity from extensive factorization of the joint posterior distribution using a mixture-of-product-of-experts approach (cf. Section). However, this advantage comes at the cost of limited applicability to datasets with a high number of modalities, where the MoPoE's quadratic factorization complexity becomes a constraint. In contrast, MAC-WAE scales linearly with the number of

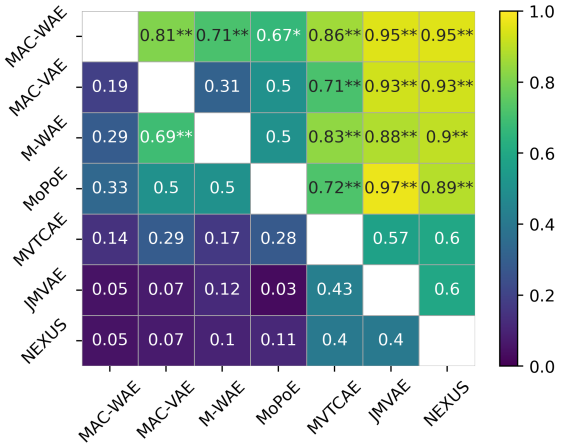


Figure 4: Heatmap depicting the relative number of outcomes (summarized in Supplementary Table 4), where the model in the given row outperformed the model in the given column. An asterisk indicates $p < 0.05$, and a double asterisk denotes $p < 0.01$ (cf. Supplementary Materials).

modalities (cf. Algorithm 1) adding an additional advantage to our method.

Known limitations Multimodal generative models may be evaluated with respect to qualities other than those we focused on in our experiments, such as probabilistic calibration, generative coherence, generative diversity, *etc*. While interesting, we viewed these as secondary objectives beyond the scope of our present empirical research investigation outlined in RQ1 and RQ2.

Future Work

The proposed model is intentionally simple, comprising a single neural encoder and a set of modality-specific neural decoders as its only trainable components; thus it bears the potential for further extensions and modifications, which we discuss in the following paragraphs.

Encoding architecture The previous works on arbitrary conditioning used encoders that were specifically designed to be robust against missing inputs. The architectures such as DeepSet (Zaheer et al. 2017), PointNet (Qi et al. 2017)), or graph attention layers (Veličković et al. 2017) can generically handle variable input structures. While introducing architectural complexity, compared to the parsimonious solution that we used in our experiments, these architectures may further enhance the performance of the model.

Regularization metrics The Wasserstein regularization penalty can be computed via several algorithms other than the MMD estimated by U-statistic. These include the following: (i) alternative MMD estimators (Yamada et al. 2018); (ii) a GAN penalty (Arjovsky, Chintala, and Bottou 2017; Tolstikhin et al. 2018) ; and (iii) sliced Wasserstein losses (Kolouri et al. 2018; Wu et al. 2019).

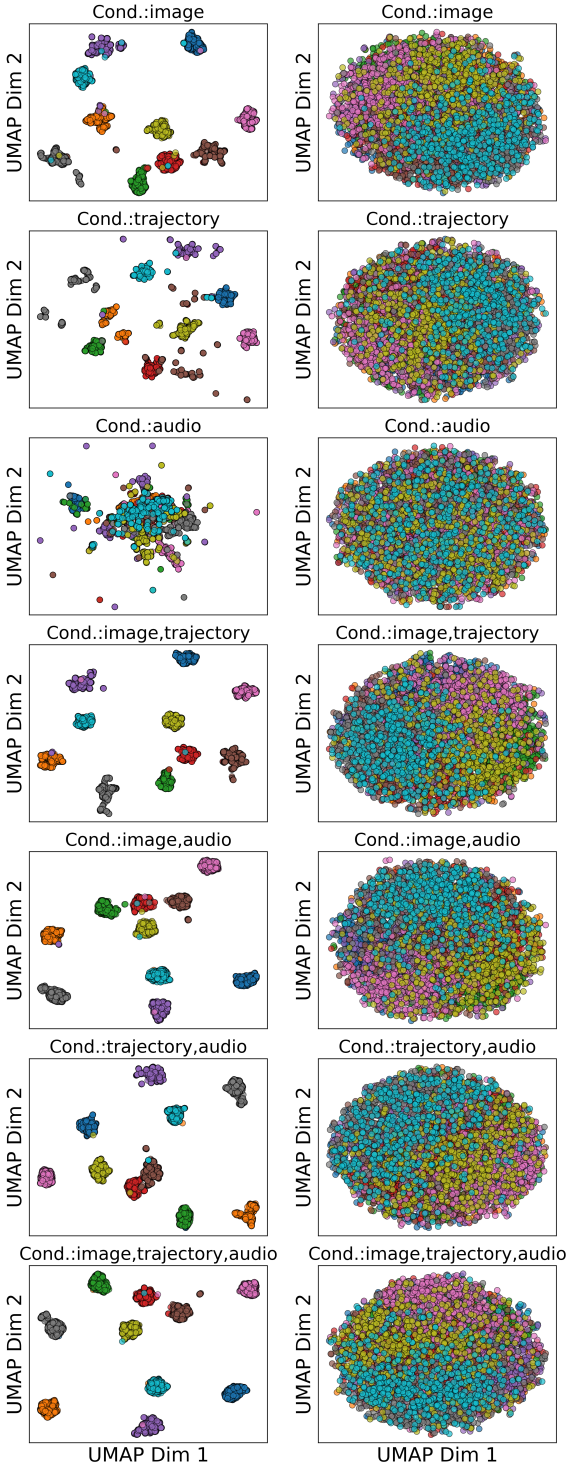


Figure 5: UMAP-processed latent codes of the MHD test samples produced by the MAC-WAE and MAC-VAE, conditioned on a subset of modalities (point colors indicate the label of the sample). The results reveal that KL-divergence (right) but not Wasserstein regularization (left) forces the latent embeddings to concentrate close to each other despite coming from diverse inputs.

References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a broken ELBO. In *International conference on machine learning*, 159–168. PMLR.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Belghazi, M.; Oquab, M.; and Lopez-Paz, D. 2019. Learning about an exponential amount of conditional distributions. *Advances in Neural Information Processing Systems*, 32.

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

Benoudjit, N.; and Verleysen, M. 2003. On the kernel widths in radial-basis function networks. *Neural Processing Letters*, 18: 139–154.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Ferguson, T. S.; Shapley, L.; and MacQueen, J. 2005. U-statistics. *University of California-Los Angeles*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

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

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. β-Vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800.

Hoffman, M. D.; and Johnson, M. J. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34: 12194–12207.

Ivanov, O.; Figurnov, M.; and Vetrov, D. 2018. Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*.

Javaloy, A.; Meghdadi, M.; and Valera, I. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *International Conference on Machine Learning*, 9938–9964. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR 2021)*.

Kolouri, S.; Pope, P. E.; Martin, C. E.; and Rohde, G. K. 2018. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Korthals, T.; Rudolph, D.; Leitner, J.; Hesse, M.; and Rückert, U. 2019. Multi-modal generative models for learning epistemic active sensing. In *2019 International Conference on Robotics and Automation (ICRA)*, 3319–3325. IEEE.

Li, Y.; Akbar, S.; and Oliva, J. B. 2020. Flow Models for Arbitrary Conditional Likelihoods. In *International Conference on Machine Learning*, 5831–5841. PMLR.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ma, C.; Tschiatschek, S.; Palla, K.; Hernandez-Lobato, J. M.; Nowozin, S.; and Zhang, C. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *International Conference on Machine Learning*, 4234–4243. PMLR.

Ma, C.; Tschiatschek, S.; Turner, R.; Hernández-Lobato, J. M.; and Zhang, C. 2020. VAEM: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33: 11237–11247.

Mahajan, S.; Botschen, T.; Gurevych, I.; and Roth, S. 2019. Joint wasserstein autoencoders for aligning multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Nazabal, A.; Olmos, P. M.; Ghahramani, Z.; and Valera, I. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107: 107501.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 4. Granada.

Noriega, L. 2005. Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, 4(5): 444.

Palumbo, E.; Daunhawer, I.; and Vogt, J. E. 2023. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. In *The Eleventh International Conference on Learning Representations*.

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

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Rubenstein, P. K.; Schoelkopf, B.; and Tolstikhin, I. 2018. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*.

Rüschendorf, L. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1): 117–129.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8247–8255.

Senellart, A.; Chadebec, C.; and Allassonnière, S. 2023. Improving Multimodal Joint Variational Autoencoders through Normalizing Flows and Correlation Analysis. *arXiv preprint arXiv:2305.11832*.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Smola, A. J.; Gretton, A.; and Borgwardt, K. 2006. Maximum mean discrepancy. In *13th international conference, ICONIP*, 3–6.

Strauss, R.; and Oliva, J. B. 2021. Arbitrary conditional distributions with energy. *Advances in Neural Information Processing Systems*, 34: 752–763.

Strauss, R.; and Oliva, J. B. 2022. Posterior matching for arbitrary conditioning. *Advances in Neural Information Processing Systems*, 35: 18088–18099.

Sutter, T.; Meng, Y.; Agostini, A.; Chopard, D.; Fortin, N.; Vogt, J.; Shahbaba, B.; and Mandt, S. 2024. Unity by Diversity: Improved Representation Learning for Multimodal VAEs. *Advances in Neural Information Processing Systems*, 37: 74262–74297.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized Multimodal ELBO. In *International Conference on Learning Representations (ICLR 2021)*.

Suzuki, M.; and Matsuo, Y. 2022. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6): 261–278.

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

Tian, Y.; and Engel, J. 2019. Latent translation: Crossing modalities by bridging generative models. *arXiv preprint arXiv:1902.08261*.

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B. 2018. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*.

Tomczak, J.; and Welling, M. 2018. VAE with a VampPrior. In *International conference on artificial intelligence and statistics*, 1214–1223. PMLR.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2017. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*.

Vasco, M.; Melo, F. S.; and Paiva, A. 2020. MH-VAE: a human-inspired deep hierarchical generative model for multimodal representation learning. *arXiv preprint arXiv:2006.02991*.

Vasco, M.; Yin, H.; Melo, F. S.; and Paiva, A. 2022. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146: 238–255.

Vedantam, R.; Fischer, I.; Huang, J.; and Murphy, K. 2017. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, J.; Huang, Z.; Acharya, D.; Li, W.; Thoma, J.; Paudel, D. P.; and Gool, L. V. 2019. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3713–3722.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.

Yamada, M.; Wu, D.; Tsai, Y.-H. H.; Takeuchi, I.; Salakhutdinov, R.; and Fukumizu, K. 2018. Post selection inference with incomplete maximum mean discrepancy estimator. *arXiv preprint arXiv:1802.06226*.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. *Advances in neural information processing systems*, 30.

Żelaszczyk, M.; and Mańdziuk, J. 2023. Cross-modal text and visual generation: A systematic review. Part 1: Image to text. *Information Fusion*, 93: 302–329.