

# Multimodal Variational Autoencoder: A Barycentric View

Peijie Qiu<sup>1</sup>, Wenhui Zhu<sup>2</sup>, Sayantan Kumar<sup>1</sup>, Xiwen Chen<sup>3</sup>, Jin Yang<sup>1</sup>, Xiaotong Sun<sup>4</sup>, Abolfazl Razi<sup>3</sup>, Yalin Wang<sup>2</sup>, Aristeidis Sotiras<sup>1\*</sup>

<sup>1</sup> Washington University in St. Louis

<sup>2</sup> Arizona State University

<sup>3</sup> Clemson University

<sup>4</sup> University of Arkansas

{peijie.qiu, sayantan.kumar, yang.jin, aristeidis.sotiras}@wustl.edu

{wzhu59, ylwang}@asu.edu, {xiwenc, arazi}@clemson.edu, xs018@uark.edu

## Abstract

Multiple signal modalities, such as vision and sounds, are naturally present in real-world phenomena. Recently, there has been growing interest in learning generative models, in particular variational autoencoder (VAE), for multimodal representation learning especially in the case of missing modalities. The primary goal of these models is to learn a modality-invariant and modality-specific representation that characterizes information across multiple modalities. Previous attempts at multimodal VAEs approach this mainly through the lens of experts, aggregating unimodal inference distributions with a product of experts (PoE), a mixture of experts (MoE), or a combination of both. In this paper, we provide an alternative generic and theoretical formulation of multimodal VAE through the lens of barycenter. We first show that PoE and MoE are specific instances of barycenters, derived by minimizing the asymmetric weighted KL divergence to unimodal inference distributions. Our novel formulation extends these two barycenters to a more flexible choice by considering different types of divergences. In particular, we explore the Wasserstein barycenter defined by the 2-Wasserstein distance, which better preserves the geometry of unimodal distributions by capturing both modality-specific and modality-invariant representations compared to KL divergence. Empirical studies on three multimodal benchmarks demonstrated the effectiveness of the proposed method.

## Introduction

Multiple data types are naturally present together to characterize the same underlying phenomena in the real world. Multimodal representation learning is thus of interest across various fields, including computer vision, natural language processing, and the biomedical domain. However, understanding and interrelating different modalities is a challenging task due to the laboriousness of human annotations and the absence of certain modalities in practice. These two factors pose a significant challenge to the application of unimodal and discriminative (supervised) representation learning methods to the multimodal case (see *e.g.*, Karpathy and Fei-Fei 2015; Pham et al. 2019; Lin et al. 2023).

Therefore, we focus on the generative models for representation learning, which are typically considered

as unsupervised, such as generative adversarial networks (GANs; Goodfellow et al. 2014) and variational autoencoders (VAEs; Kingma and Welling 2013). In particular, we focus on VAEs for multimodal representation learning since VAEs are graphical probabilistic models capable of learning an explicit latent distribution, which has the potential to directly learn the joint distributions of multiple modalities (Suzuki, Nakayama, and Matsuo 2016; Baltrušaitis, Ahuja, and Morency 2018). Despite their nice probabilistic properties and the success in unimodal applications, the direct translation of VAEs to the multimodal case (*e.g.*, feeding the multimodal data to VAEs) is challenging, as they struggle with handling missing modalities and performing cross-modal generations. Therefore, the design of multimodal VAEs seeks to form a modality-invariant and modality-specific latent representation by learning a joint latent distribution (so-called joint posterior) to aggregate the information from different modalities (Ngiam et al. 2011; Suzuki, Nakayama, and Matsuo 2016; Baltrušaitis, Ahuja, and Morency 2018). The modality-specific and modality-invariant formulation naturally enables a cross-modal generation (Shi et al. 2019). In addition, it can also handle missing modalities by directly sampling the learned joint posterior.

The core objective of multimodal VAEs then revolves around how to approximate the joint posterior by aggregating the unimodal posterior, also known as unimodal inference distribution in VAEs. This typically involves finding a proper aggregation function. However, such aggregation functions are challenging to identify due to the intractability of the true joint posterior. Previous explorations of multimodal VAEs addressed this challenge mainly through the lens of experts in statistics by aggregating unimodal inference distributions with a product of experts (PoE; Wu and Goodman 2018), a mixture of experts (MoE; Shi et al. 2019), or a combination of both (MoPoE; Sutter, Daunhawer, and Vogt 2021). Although empirical studies have shown their success for multimodal VAEs, theoretical analysis of their properties is still insufficient.

In this paper, we provide a theoretical view of previous multimodal VAEs in a unified way through the lens of barycenter. The barycentric distribution is the mean distribution of a set of distributions, defined by minimizing the weighted sum of divergences to these distributions. Interestingly, we discovered that the distributions aggregated

\*Corresponding author

by PoE and MoE are barycenters by optimizing the reverse and forward Kullback-Leibler (KL) divergence, respectively. This directly provides an information-theoretic view of PoE and MoE, which reveals their intrinsic properties: PoE is zero-forcing (*i.e.*, pushing the joint posterior biased towards certain modalities), while MoE is mass-covering (*i.e.*, balancing all modalities). However, the KL divergence does not define a metric space for probability measures, as it is asymmetric and unbounded. This motivates us to explore other divergence measures that are defined in metric space. In particular, we explored the Wasserstein barycenter (Agueh and Carlier 2011) by optimizing the squared 2-Wasserstein distance, as it preserves the geometry of unimodal inference distributions in a geodesic space (whereas KL divergence focuses on pointwise differences). Leveraging the intricate geometry of the Wasserstein distance (Peyré, Cuturi et al. 2019), the Wasserstein barycenter serves as the Fréchet means (see *e.g.*, Grove and Karcher 1973) within the space of probability measures.

In summary, our contributions are threefold: **i)** We introduce a novel and unified formulation for multimodal VAEs, where the aggregation of unimodal inference distributions is framed as solving the barycenter problem that minimizes certain divergence measures. This approach offers a theoretical framework to analyze intrinsic properties and enables a more flexible selection of aggregation functions for multimodal VAEs. **ii)** We propose  $\mathcal{WB}$ -VAE, a novel multimodal VAE for representation learning that leverages the Wasserstein barycenter to aggregate unimodal inference distributions. **iii)** Experiments on three benchmark datasets demonstrated the effectiveness of the proposed method compared to other state-of-the-art methods.

## Background and Related Work

### Multimodal VAEs

Prior multimodal VAEs can be roughly divided into two main categories: coordinated models and joint models. The former only learns the inference distributions from a single modality, while the latter learns the joint inference distributions across all modalities (Baltrušaitis, Ahuja, and Morency 2018; Suzuki and Matsuo 2022). Accordingly, coordinated models (Higgins et al. 2017; Schonfeld et al. 2019; Korthals et al. 2019) strive to generate consistent inference results across all modalities. Although they can perform cross-modal generation, they may not effectively handle missing modalities as in joint models (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020). This is because they do not model the joint inference distribution of all modalities as in joint models.

Here, we focus on joint models that can be applied to a wider spectrum of applications. Although there are some joint models that can handle missing modalities via a surrogate unimodal inference model (Vedantam et al. 2017; Korthals et al. 2019), they typically face scalability issues. Hence, we consider joint models that can directly learn the joint inference distributions by aggregating unimodal inference distributions through an aggregation function. Following this vein, Wu and Goodman (2018) proposed an PoE-

VAE (*a.k.a.*, MVAE) by aggregating the unimodal distributions with a product of experts. Despite resulting in a sharper joint distribution, PoE-VAE is prone to focus on certain modalities while neglecting others. To mitigate this issue, Shi et al. (2019) proposed an MoE-VAE (*a.k.a.*, MM-VAE) by leveraging a mixture of experts. However, MoE-VAE does not produce a joint distribution that is sharper than any other expert: the precision of the joint inference distribution may not increase as the number of modalities increases. To take advantage of both PoE and MoE, Sutter, Daunhawer, and Vogt (2021) proposed a generalized MoPoE-VAE, which first applies PoE and then MoE to all possible subsets of modalities. However, the previous attempts at joint models are limited to the perspective of experts in statistics.

Although there are other multimodal VAEs (Palumbo, Daunhawer, and Vogt 2023; Hirt et al. 2024; Yuan et al. 2024), their focus is not on new aggregation functions. Instead, they are considered variants of PoE-VAE and MoE-VAE. In this paper, we provide a unified framework for aggregation functions from a barycentric view. In contrast to previous works that combined unimodal distribution aggregation with model parameter optimization (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020, 2021), our barycentric formulation decouples these two steps. This enables a more flexible choice of barycenters for aggregating unimodal inference distributions (*e.g.*, the Wasserstein barycenter, which we explore in this paper).

### Optimal Transport and Wasserstein distance

Optimal transport (OT) seeks to find a transport map to move the mass from one distribution to another while minimizing the transport cost. Here, we consider Kantorovich’s dual OT formulation (Kantorovich 1942) instead of Monge’s primal formulation (Monge 1781), as Monge’s formulation is not symmetric.  $P \sim \mathcal{P}(\mathcal{X})$  and  $Q \sim \mathcal{P}(\mathcal{Y})$ , with  $\mathcal{P}(\mathcal{X})$  and  $\mathcal{P}(\mathcal{Y})$  being the respective sets of probability distributions on them, Kantorovich’s OT formulation is defined as

$$\inf_{\pi \in \Pi(P, Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y),$$

where  $c : \mathcal{X} \times \mathcal{Y}$  is a cost function. The infimum is taken over the set of all transport plans  $\pi \in \Pi(P, Q)$ , *i.e.*, joint distributions on  $\mathcal{X} \times \mathcal{Y}$  with marginals  $P$  and  $Q$ .

The  $p$ -Wasserstein distance is then the  $p$ -th root of the infimum of Kantorovich’s OT formulation for a cost function  $c(x, y) = |x - y|^p$ :

$$\mathcal{W}_p(P, Q) = \inf_{\pi \in \Pi(P, Q)} \left( \int_{\mathcal{X} \times \mathcal{Y}} |x - y|^p d\pi(x, y) \right)^{1/p},$$

with  $p = 1$  being an earth mover’s distance that is commonly used in many generative adversarial networks (see *e.g.*, Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Miyato et al. 2018). In contrast, we focus on the 2-Wasserstein distance for deriving the Wasserstein barycenter in this paper, as its quadratic form allows for an analytic solution in the case of Gaussian distributions. For two Gaussian distributions  $\mathcal{N}(\mu_1, \Sigma_1)$  and  $\mathcal{N}(\mu_2, \Sigma_2)$ , the squared

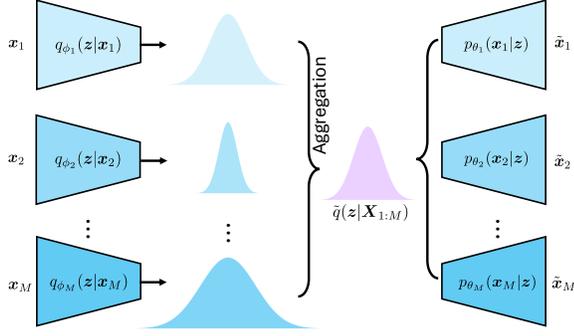


Figure 1: The overview of a multimodal VAE that takes  $M$  modalities  $\mathbf{X}_{1:M} = \{\mathbf{x}_j\}_{j=1}^M$  as input and outputs the reconstructed input modalities  $\tilde{\mathbf{X}}_{1:M} = \{\tilde{\mathbf{x}}_j\}_{j=1}^M$ . The multimodal VAE consists of  $M$  probabilistic encoders  $\{q_{\phi_j}(\mathbf{z}|\mathbf{x}_j)\}_{j=1}^M$  and decoders  $\{p_{\theta_j}(\mathbf{x}_j|\mathbf{z})\}_{j=1}^M$ .

2-Wasserstein distance between them is solved analytically (see *e.g.*, Knott and Smith 1984; Givens and Shortt 1984):

$$\mathcal{W}_2^2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_2^2 + \text{Tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2}). \quad (1)$$

## Methods

### Multimodal VAE: an Expert View

Without loss of generality, we consider a dataset  $\{\mathbf{X}_{1:M}^{(i)}\}_{i=1}^N$  containing  $N$  number of independent and identically distributed (i.i.d.) samples, each of which consists of  $M$  modalities:  $\mathbf{X}_{1:M}^{(i)} = \{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_M^{(i)}\}$ . Assuming the multimodal data can be generated by some random process involving a joint latent variable  $\mathbf{z}$ , the objective of a multimodal VAE is to maximize the log-likelihood of data over all  $M$  modalities, given i.i.d. condition:

$$\log p_{\theta}(\mathbf{X}_{1:M}^{(i)}) = D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{X}_{1:M}^{(i)})||p_{\theta}(\mathbf{z}|\mathbf{X}_{1:M}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{X}_{1:M}^{(i)}), \quad (2)$$

where  $q_{\phi}(\mathbf{z}|\mathbf{X}_{1:M}^{(i)})$  is the approximate posterior parameterized by deep neural networks (*i.e.*, the probabilistic encoders in VAEs), as the true posterior is intractable in practice. Since the KL divergence of the approximate from the true posterior (*i.e.*, first RHS term in Eq. (2)) is non-negative, we instead maximize the evidence lower bound (ELBO)  $\mathcal{L}(\theta, \psi; \mathbf{X}_{1:M}^{(i)})$  as follows:

$$\mathcal{L}(\theta, \phi; \mathbf{X}_{1:M}^{(i)}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{X}_{1:M}^{(i)})}[\log p_{\theta}(\mathbf{X}_{1:M}^{(i)}|\mathbf{z}) - D_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{X}_{1:M}^{(i)})||p_{\theta}(\mathbf{z}))], \quad (3)$$

where  $\{q_{\phi_m}(\mathbf{z}|\mathbf{X})\}_{m=1}^M$  and  $\{p_{\theta_m}(\mathbf{X}|\mathbf{z})\}_{m=1}^M$  are the  $M$  probabilistic encoders and decoders, respectively. For notation brevity, we will omit the sample index ( $i$ ) hereafter. An overview of the multimodal VAE is shown in Fig. 1. However, in a multimodal scenario, maximizing the above

ELBO objective requires the knowledge of the true joint posterior  $p_{\theta}(\mathbf{z}|\mathbf{X}_{1:M})$ , which is unknown in practice. To tackle this issue, previous explorations of multimodal VAEs approximate the true joint posterior by aggregating the unimodal inference distributions with a proper function  $f_{\text{aggr}}(\cdot)$ :  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M}) = f_{\text{aggr}}(\{q_{\phi_m}\}_{m=1}^M)$ , where  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M})$  denotes the approximate joint posterior. Some popular choices of  $f_{\text{aggr}}(\cdot)$  are PoE (Wu and Goodman 2018), MoE (Shi et al. 2019), or a combination of both (MoPoE; Sutter, Daunhawer, and Vogt 2021). Formally, the approximate joint posterior by PoE and MoE can be summarized as

$$\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M}) = \begin{cases} \frac{1}{Z} \prod_{m=1}^M q_{\phi_m}(\mathbf{z}|\mathbf{x}_m), & \text{PoE,} \\ \frac{1}{M} \sum_{m=1}^M q_{\phi_m}(\mathbf{z}|\mathbf{x}_m), & \text{MoE,} \end{cases}$$

where  $Z$  is the normalizer function that ensures the approximate posterior by PoE is a valid probability measure.

### Multimodal VAE: a Barycentric View

The barycenter of distribution is defined as a central distribution of a set of distributions that minimizes the sum of divergences to all other distributions in the set. For a set of probability distributions  $\{P_1, \dots, P_M\}$  with associated weights  $\{\lambda_1, \dots, \lambda_M\}$ , the barycenter minimizes the weighted sum of some divergences  $d(\cdot, \cdot)$  from the barycenter distribution  $P_{\mathcal{B}}$  to each of the given distributions:

$$P_{\mathcal{B}} = \arg \min_P \sum_{m=1}^M \lambda_m d(P_m, P), \quad \sum_{m=1}^M \lambda_m = 1.$$

**Lemma 1.** *In the context of multimodal VAE, we seek to find a barycenter  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M})$  that can aggregate the unimodal inference distributions  $\{q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)\}_{m=1}^M$  to approximate the true joint posterior  $p_{\theta}(\mathbf{z}|\mathbf{X}_{1:M})$ :*

$$\tilde{q} = \arg \min_q \sum_{m=1}^M \lambda_m d(q_{\phi_m}, q), \quad \sum_{m=1}^M \lambda_m = 1. \quad (4)$$

Note that, for notation brevity, we abbreviate  $q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)$  and  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M})$  as  $q_{\phi_m}$  and  $\tilde{q}$ , respectively. Instead of directly minimizing the divergence between  $q_{\phi}(\mathbf{z}|\mathbf{X}_{1:M})$  and  $p_{\theta}(\mathbf{z})$  over trainable parameters  $\phi = \{\phi_1, \dots, \phi_M\}$  as formulated in Eq. (3) and prior multimodal VAEs (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020, 2021), Lemma 1 suggests that this involves a bilevel optimization. For the lower-level optimization (*i.e.*, Eq. (4)), we determine a barycenter  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M})$ , which is equivalent to applying an aggregation function  $f_{\text{aggr}}$  to combine the unimodal inference distributions. We then push  $\tilde{q}(\mathbf{z}|\mathbf{X}_{1:M})$  towards  $p_{\theta}(\mathbf{z})$  by minimizing their divergence over trainable parameters  $\phi = \{\phi_m\}_{m=1}^M$  (upper-level optimization; Eq. (3)). At first glance, this formulation is counterintuitive, as it complicates the formulation and optimization, whereas in-depth analysis reveals its theoretically intriguing properties.

**Proposition 1.** *For any divergence measure  $d(q_{\phi_m}, \cdot)$  that is convex on  $q_{\phi_m}$ , the resulting barycenter by minimizing*

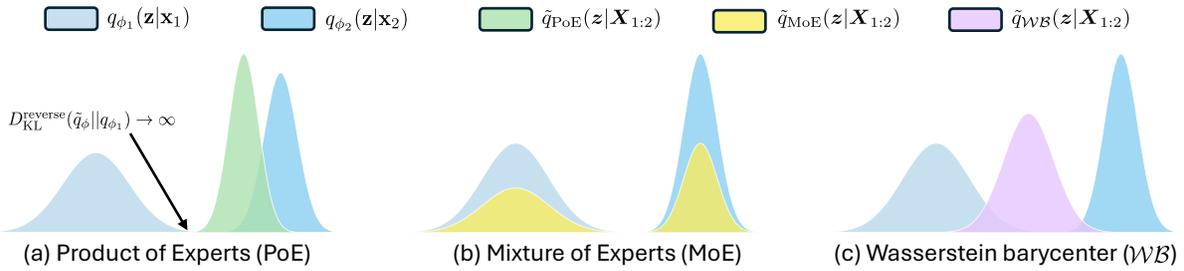


Figure 2: Comparison of methods for aggregating the unimodal inference distributions ( $\{q_{\phi_j}\}_{j=1}^M$ ) to approximate the joint posterior ( $\tilde{q}_\phi$ ): (a) PoE, (b) MoE, and (c) the proposed Wasserstein barycenter. In this illustrative example, we use two 1-dimensional Gaussian modalities ( $M = 2$ ) for a proof of concept.

Eq. (4) guarantees a valid ELBO on the marginal log-likelihood  $p_\theta(\mathbf{X}_{1:M})$  and a scalable inference. This is because of Jensen’s inequality:

$$d\left(\sum_{m=1}^M \lambda_m q_{\phi_m}, q\right) \leq \sum_{m=1}^M \lambda_m d(q_{\phi_m}, q) \quad (5)$$

For a complete proof of Proposition 1, please see **Appendix A.1**. The LHS in Eq. (5) defines a scalable inference, as the naive implementation on the RHS requires  $2^M$  inference networks to handle arbitrary combination of input modalities. Although Proposition 1 has been considered in some prior works from different perspectives (Shi et al. 2019; Sutter, Daunhawer, and Vogt 2021), they are limited to the case of KL divergence (see Theorem 1). In contrast, our barycentric view extends them to a more general case whenever  $d(q_{\phi_m}, q)$  is convex to  $q_{\phi_m}$ , which enables it to analyze the properties of a more flexible choice of divergence measures (e.g.,  $f$ -divergence, 2-Wasserstein distance, Gromov-Wasserstein distance, etc).

**Theorem 1.** Considering KL divergence  $D_{KL}(\cdot||\cdot)$  as the divergence measure  $d(\cdot, \cdot)$ , PoE and MoE are the barycenters yielded by optimizing the reverse and forward KL divergence, respectively:

$$\tilde{q}_{PoE} = \arg \min_q \frac{1}{Z} \sum_{m=1}^M D_{KL}^{reverse}(q||q_{\phi_m}),$$

$$\tilde{q}_{MoE} = \arg \min_q \frac{1}{M} \sum_{m=1}^M D_{KL}^{forward}(q_{\phi_m}||q).$$

The proof of Theorem 1 is in **Appendix A.2**. In information theory, it is customary to define KL divergence as relative entropy (due to its asymmetry), with the form used in PoE and MoE in Theorem 1 being the exclusive (reverse) and inclusive (forward) KL divergence (Cover 1999; Murphy 2012). Theorem 1 immediately provides an information-theoretic view of PoE and MoE: they are two variants resulting from the inherent asymmetry of KL divergence. This provides us with an information-theoretic tool to analyze the properties of PoE and MoE in multimodal VAE.

**Remark 1.** PoE is zero-forcing, encouraging  $\tilde{q}(z|\mathbf{X}_{1:M})$  to be zero where  $q_{\phi_m}(z|\mathbf{x}_m)$  is zero, which makes it bi-

ased towards certain modalities. In contrast, MoE is mass-covering, ensuring that there is mass under  $\tilde{q}(z|\mathbf{X}_{1:M})$  wherever there is mass under  $q_{\phi_m}(z|\mathbf{x}_m)$ .

Remark 1 is due to the intrinsic properties of forward and reverse KL divergence (Minka et al. 2005; Turner and Sahani 2011). Though it is well known that PoE results in a sharper distribution that concentrates on one of the modalities, whereas MoE does not produce a distribution sharper than any individual expert due to the nature of the mixture, Remark 1 provides an information-theoretic interpretation. We demonstrate this by considering an example with two modalities, as shown in Fig. 2. When there is zero mass under  $q_{\phi_1}$  and nonzero mass under  $\tilde{q}_{PoE}$ , the reverse KL divergence is almost infinity:  $D_{KL}^{reverse}(\tilde{q}_{PoE}||q_{\phi_1}) \rightarrow \infty$ , which pushes  $\tilde{q}_{PoE}$  toward  $q_{\phi_2}$  (see Fig. 2a). In contrast, since the forward KL divergence penalizes  $\log q_{\phi_m}(z|\mathbf{x}_m) - \log \tilde{q}(z|\mathbf{X}_{1:M})$ , it ensures that  $\tilde{q}$  has mass covered wherever this is mass under  $q_{\phi_m}$  (see Fig. 2b).

However, the forward and reverse KL divergence does not define a metric space for probability measures because it is asymmetric and unbounded. One notable example is that solving Eq. (4) does not guarantee a valid probability measure in the case of PoE (see **Appendix A.2**). This motivates us to find a barycenter defined in the probability metric space. Below, we explore the barycenter defined in the 2-Wasserstein space, known as the Wasserstein barycenter.

### Multimodal VAE from Wasserstein Barycenter

Here, we provide a roadmap to derive the proposed Wasserstein barycenter VAE (WB-VAE) for multimodal representation learning. Following the convention in Eq. (4), Wasserstein barycenter (WB) is defined by minimizing the sum of the squared 2-Wasserstein distance  $\mathcal{W}_2^2(\cdot, \cdot)$ . Since the 2-Wasserstein distance is symmetric, the order of distributions in  $\mathcal{W}_2(\cdot, \cdot)$  does not matter. In the context of multimodal VAE, the approximate posterior resulting from optimizing the squared 2-Wasserstein distance is

$$\tilde{q}_{WB} = \arg \min_q \sum_{m=1}^M \lambda_m \mathcal{W}_2^2(q_{\phi_m}, q), \quad \sum_{m=1}^M \lambda_m = 1.$$

Unlike the KL divergence used in the case of PoE and MoE, which focuses on pointwise differences, the 2-Wasserstein

distance better preserves the geometry of the unimodal inference distributions. Accordingly, interpolating in the Wasserstein space (*i.e.*, a geodesic space) can have a meaningful transition from unimodal distributions to the joint posterior, especially when the unimodal distributions have different shapes or supports (Ambrosio, Gigli, and Savaré 2008). Therefore, different choices of weights associated with unimodal distributions (*i.e.*,  $\{\lambda_1, \dots, \lambda_M\}$ ) may lead to a joint posterior that maintains diverse shapes and structures of unimodal distributions. However, in the context of multimodal VAEs, it is challenging to determine  $\{\lambda_1, \dots, \lambda_M\}$ , as we only have the marginal unimodal distributions. Similar to the case of PoE and MoE, it is typically safe to set  $\lambda_m = 1/M, \forall m$ .

**Bures-Wasserstein barycenter.** Wasserstein barycenter typically incurs the significant computational cost associated with the 2-Wasserstein distance. However, in the case of Gaussian distributions, as are typically assumed in VAEs, the Gaussian Wasserstein barycenter (*i.e.*, the so-called Bures-Wasserstein barycenter (Agueh and Carlier 2011)) can be obtained by solving a fixed-point equation (Knott and Smith 1994; Agueh and Carlier 2011).

Considering the unimodal inference distributions  $\{q_{\phi_m}\}_{m=1}^M$  are  $d$ -dimensional multivariate Gaussian  $\{\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\}_{m=1}^M$ , with  $\boldsymbol{\mu}_m \in \mathbb{R}^d$  and  $\boldsymbol{\Sigma}_m \in \mathbb{R}^{d \times d}$  being the associated mean and covariance of  $q_{\phi_m}$ , the resulting Bures-Wasserstein barycenter turns out to be Gaussian-distributed, *i.e.*,  $\tilde{q}_{\mathcal{WB}}(\mathbf{z}|\mathbf{X}_{1:M}) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ :

$$\tilde{\boldsymbol{\mu}} = \sum_{m=1}^M \lambda_m \boldsymbol{\mu}_m, \quad \tilde{\boldsymbol{\Sigma}} = \sum_{m=1}^M \lambda_m (\tilde{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\Sigma}_m \tilde{\boldsymbol{\Sigma}}^{1/2})^{1/2}, \quad (6)$$

where the covariance  $\tilde{\boldsymbol{\Sigma}}$  is obtained by solving the fixed-point iteration. However, Eq. (6) can be further simplified by considering  $q_{\phi_m}(\mathbf{z}|\mathbf{x}_m)$  an isotropic Gaussian with a diagonal covariance  $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2 \mathbf{I})$  with  $\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m \in \mathbb{R}^d$  and  $\mathbf{I} \in \mathbb{R}^{d \times d}$ . This is typically assumed in most VAEs (Kingma and Welling 2013).

**Remark 2.** *In the isotropic Gaussian case, Eq. (6) can be solved analytically dimension by dimension:*

$$\tilde{\boldsymbol{\mu}} = \sum_{m=1}^M \lambda_m \boldsymbol{\mu}_m, \quad \tilde{\boldsymbol{\sigma}} = \sum_{m=1}^M \lambda_m \boldsymbol{\sigma}_m. \quad (7)$$

Remark 2 is because the optimal transport map from one Gaussian to another is a linear map (Knott and Smith 1994; Agueh and Carlier 2011), with which the squared 2-Wasserstein distance can be solved analytically (for details, please see **Appendix A.3**). As suggested by Lemma 1, the Bures-Wasserstein barycenter can be viewed as minimizing the 2-Wasserstein distance to a mixture of distributions.

**Mixture of Wasserstein barycenter.** The approximate joint distribution derived from solving the Wasserstein barycenter strikes a balance between zero-forcing (bias) and mass-covering (variance), resulting in a distribution that is sharper than half of the unimodal inference distributions (see Fig. 2c). However, there is an inherent trade-off between

zero-forcing and mass-covering (Murphy 2012). Similar to MoPoE-VAE (Sutter, Daunhawer, and Vogt 2021), we consider a variant of  $\mathcal{WB}$ -VAE by constructing a mixture of Wasserstein barycenter, termed  $\mathcal{MWB}$ -VAE.

**Remark 3.** *The mixture of Wasserstein barycenter with unimodal inference distributions is still a barycenter. Considering the powerset of  $M$  modalities  $\mathcal{P}_M(\mathbf{X})$ , which consists of  $2^M$  different combinations, the mixture of Wasserstein barycenter is given as*

$$\begin{aligned} \tilde{q}_{\mathcal{MWB}} &= \arg \min_q \sum_{\mathbf{X}_k \in \mathcal{P}_M(\mathbf{X})} \lambda_k D_{\text{KL}}(\tilde{q}_{\mathcal{WB}} || q) \\ \text{subject to } \tilde{q}_{\mathcal{WB}} &= \arg \min_q \sum_{\mathbf{x}_j \in \mathbf{X}_k} \lambda_j \mathcal{W}_2^2(q_{\phi_j}, q) \end{aligned}$$

Though this is a bilevel optimization problem, the solution is analytical since both the lower-level and upper-level optimization problems can be solved analytically. The solution is also optimal due to the convexity of both forward KL divergence and 2-Wasserstein distance. By applying the same mechanism, we can also derive MoPoE (Sutter, Daunhawer, and Vogt 2021) as a barycenter, whereas the solution is not guaranteed to be optimal since the solution to the lower-level (PoE) case is not a global optimum in general.

## Experiments

**Dataset.** We conducted comparative experiments on three multimodal benchmark datasets: i) PolyMNIST with five simplified modalities, ii) the trimodal MNIST-SVHN-TEXT, and iii) the challenging bimodal CelebA dataset. PolyMNIST was generated by combining each MNIST digit (LeCun and Cortes 2010) with  $28 \times 28$  random crops from five distinct background images, as described in (Sutter, Daunhawer, and Vogt 2021). The MNIST-SVHN-TEXT dataset was introduced by (Sutter, Daunhawer, and Vogt 2020), which consists of three modalities: MNIST digit (LeCun and Cortes 2010), text, and SVHN (Netzer et al. 2011). The MNIST digit and text are two clean modalities, whereas SVHN is comprised of noisy images. Following (Sutter, Daunhawer, and Vogt 2021), 20 triples were generated per set using a many-to-many mapping. The bimodal CelebA includes human face images as well as text describing the face attributes (Liu et al. 2015). This dataset is challenging because the text modality focuses on the attributes present in a face image. If an attribute is absent, it is omitted from the corresponding text (Sutter, Daunhawer, and Vogt 2020).

**Baseline methods.** We compared the proposed method to three state-of-the-art multimodal VAEs, including PoE-VAE (Wu and Goodman 2018), MoE-VAE (Shi et al. 2019), and MoPoE-VAE (Sutter, Daunhawer, and Vogt 2021).

**Evaluation metric.** Following previous literature in Wu and Goodman (2018); Shi et al. (2019); Sutter, Daunhawer, and Vogt (2021), several tasks were conducted to evaluate the performance of the multimodal VAEs. First, a linear classifier was used to assess the quality of the learned latent representations. Second, the coherence of generated samples was evaluated using pre-trained classifiers. Third, the approximate joint posterior was measured by calculating the log-likelihoods on the test set.

Model	M	S	T	M,S	M,T	S,T	M,S,T	Avg.
PoE-VAE	0.90±0.01	0.44±0.01	0.85±0.10	0.89±0.01	0.97±0.02	0.81±0.09	0.96±0.02	0.83
MoE-VAE	0.95±0.01	0.79±0.05	0.99±0.01	0.87±0.03	0.93±0.03	0.84±0.04	0.86±0.03	0.89
MoPoE-VAE	0.95±0.01	0.80±0.03	0.99±0.01	0.97±0.01	0.98±0.01	0.99±0.01	0.98±0.01	0.95
$\mathcal{WB}$ -VAE	0.91±0.03	0.44±0.02	1.00±0.00	0.89±0.00	0.99±0.02	0.99±0.01	0.99±0.00	0.89
$\mathcal{MWB}$ -VAE	<b>0.97±0.00</b>	<b>0.83±0.01</b>	<b>1.00±0.00</b>	<b>0.99±0.00</b>	<b>1.00±0.00</b>	<b>1.00±0.00</b>	<b>1.00±0.00</b>	<b>0.97*</b>

(a) Linear classification accuracy of latent representations

Model	M			S			T			Avg.
	S	T	S,T	M	T	M,T	M	S	M,S	
PoE-VAE	0.24	0.20	0.32	<b>0.43</b>	0.30	<b>0.75</b>	0.28	0.17	0.29	0.32
MoE-VAE	0.75	0.99	0.87	0.31	0.30	0.30	0.96	0.76	0.84	0.68
MoPoE-VAE	0.74	0.99	0.94	0.36	0.34	0.37	0.96	0.76	0.93	0.71
$\mathcal{WB}$ -VAE	0.12	0.51	0.57	0.28	<b>0.39</b>	0.53	0.52	0.18	0.57	0.41
$\mathcal{MWB}$ -VAE	<b>0.82</b>	<b>1.00</b>	<b>0.99</b>	0.36	0.35	0.39	<b>0.97</b>	<b>0.84</b>	<b>0.99</b>	<b>0.75*</b>

(b) Conditional generation coherence

Model	$\mathbf{X}$	$\mathbf{X} x_M$	$\mathbf{X} x_S$	$\mathbf{X} x_T$	$\mathbf{X} x_M, x_S$	$\mathbf{X} x_M, x_T$	$\mathbf{X} x_S, x_T$
PoE-VAE	-1790±3.3	-2090±3.8	-1895±0.2	-2133±6.9	-1825±2.6	-2050±2.6	<b>-1855±0.3</b>
MoE-VAE	-1941±5.7	<b>-1987±1.5</b>	<b>-1857±12</b>	<b>-2018±1.6</b>	-1912±7.3	-2002±1.2	-1925±7.7
MoPoE-VAE	-1819±5.7	-1991±2.9	<b>-1858±6.2</b>	-2024±2.6	-1822±5.0	<b>-1987±3.1</b>	<b>-1850±5.8</b>
$\mathcal{WB}$ -VAE	<b>-1785±7.4</b>	-2072±13	-1889±7.4	-2126±12	<b>-1814±7.5</b>	-2033±7.1	<b>-1856±4.7</b>
$\mathcal{MWB}$ -VAE	-1890±1.7	-2000±1.4	<b>-1856±3.4</b>	-2036±0.4	-1825±1.6	<b>-1988±1.4</b>	<b>-1853±2.2</b>

(c) Log-likelihoods of the joint generative model (ranking based on first three decimals)

Table 1: Quantitative results on MNIST-SVHN-TEXT in terms of (a) linear classification accuracy, (b) conditional generation coherence, and (c) log-likelihoods of the joint generative model. We evaluated all possible combinations of modalities  $\mathbf{X}_k$ . We reported the means ( $\pm$  standard deviations) over 5 runs, where the best performance is highlighted with **bold**. The abbreviations of different modalities in this table are as follows: M: MNIST; S: SVHN; T: Text.

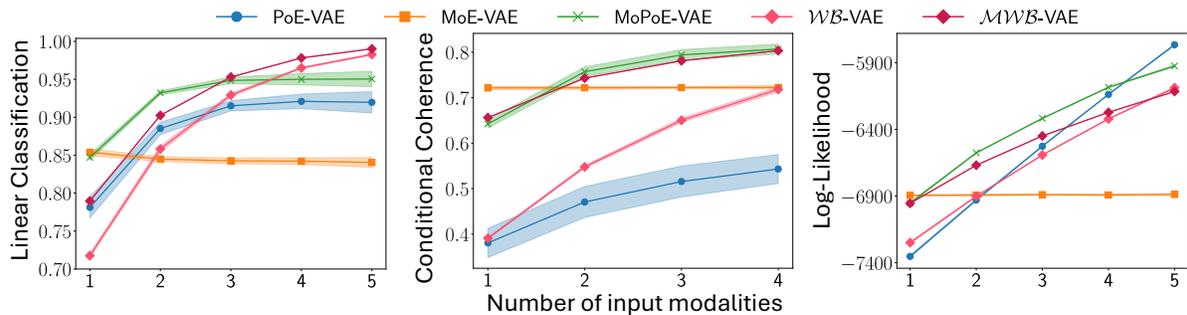


Figure 3: Quantitative results on PolyMNIST as a function of the number of input modalities, averaged over all subsets of modalities of the respective size. **Left:** Linear classification accuracy of digits given the latent representation. **Center:** Coherence of conditionally generated samples that do not include input modalities. **Right:** Log-Likelihood of all generated modalities.

**Implementation details.** For a fair comparison, we followed the experimental settings in previous literature (Shi et al. 2019; Sutter, Daunhawer, and Vogt 2021). In particular, we employed the same network architecture as in (Shi et al. 2019; Sutter, Daunhawer, and Vogt 2021). For more implementation details (*e.g.*, hyperparameter configurations), we kindly direct the readers to **Appendix B**. All experiments were performed on a Nvidia-A100 GPU with 40G memory.

## Results

**MNIST-SVHN-TEXT results.** As shown in Tables 1, the proposed  $\mathcal{MWB}$ -VAE demonstrated superior performance compared to other state-of-the-art multimodal VAEs in terms of the quality of learned latent representations and generation coherence. In addition, our  $\mathcal{WB}$ -VAE outper-

formed PoE-VAE regarding the linear classification accuracy using the learned latent representations and was on par with PoE-VAE regarding generation coherence. Although there is an inherent trade-off between generation coherence and log-likelihood, the log-likelihood of our  $\mathcal{WB}$ -VAE and  $\mathcal{MWB}$ -VAE were on par with the other state-of-the-art methods. This suggests that the proposed method can approximate the joint posterior well.

**PolyMNIST results.** The PolyMNIST dataset is unique in that it contains more than three modalities, enabling us to explore how different methods perform as the number of input modalities increases (see Fig. 3). Notably, the proposed  $\mathcal{WB}$ -VAE and  $\mathcal{MWB}$ -VAE showed an approximately linear relationship between all the performance metrics and the number of input modalities. This was particularly true for the linear classification task, where the performance of other

5 o'clock shadow, male, bags under eyes, bangs, black hair, hair, bushy eyebrows, mustache, narrow eyes, straight hair

Brown hair, heavy makeup, high cheekbones, mouth slightly open, no beard, nose, smiling

Bangs, brown hair, high cheekbones, male, mouth slightly open, no beard, oval face, smiling

Black hair, double chin, high cheekbones, , mouth slightly open, no beard, smiling

Blond hair, heavy makeup, high cheekbones, no beard, pointy nose, smiling, wavy hair, wearing lipstick

Chubby, double chin, eyeglasses, gray hair, male, no beard, receding hairline

Attractive, heavy makeup, narrow eyes, no beard, pointy nose, wearing earrings, young

big nose, black hair, bushy eyebrows, high cheekbones, male, mouth slightly open, smiling

5 o'clock shadow, attractive, black hair, eyeglasses, male, sideburns, straight hair

High cheekbones, no beard, oval face, pointy nose, smiling, wearing earrings



Figure 4: Conditionally generated images given the text on top of each column on bimodal CelebA using  $\mathcal{MWB}$ -VAE.

Model	Latent Representation			Generation		
	I	T	Joint	I $\rightarrow$ T	T $\rightarrow$ I	Avg.
PoE-VAE	0.30	0.31	0.32	0.26	0.33	0.30
MoE-VAE	0.35	0.38	0.35	0.14	0.41	0.33
MoPoE-VAE	<b>0.40</b>	0.39	0.39	0.15	<b>0.43</b>	0.35
$\mathcal{WB}$ -VAE	0.34	0.38	0.40	0.29	0.40	0.36
$\mathcal{MWB}$ -VAE	0.37	<b>0.44</b>	<b>0.44</b>	<b>0.34</b>	<b>0.43</b>	<b>0.40*</b>

Table 2: Classification accuracy based on latent representation and conditionally generated coherence on the bimodal CelebA dataset. We report the mean average precision over all attributes (I: Image; T: Text; Joint: I and T).

baseline methods was typically saturated after reaching a certain number of modalities (*e.g.*,  $M > 3$  in Fig. 3 **Left**). As a consequence,  $\mathcal{WB}$ -VAE and  $\mathcal{MWB}$ -VAE showed superior performance in terms of linear classification accuracy compared to all baseline methods, particularly when the number of input modalities increases. Similar trends were also observed in the conditional generation task (Fig. 3 **Center**), where the generation coherence of  $\mathcal{WB}$ -VAE increased as the number of input modalities increased. Although  $\mathcal{WB}$ -VAE outperformed PoE-VAE, it did not surpass MoE-VAE, but it struck the balance between them, as there is an inherent trade-off between mass-covering and zero-forcing. As a consequence,  $\mathcal{MWB}$ -VAE can easily outperform MoE-VAE and achieve similar performance as MoPoE-VAE in the conditional generation task. As suggested by Sutter, Daunhawer, and Vogt (2021), there is a trade-off between generation coherence and the log-likelihood. Consequently, the PoE-VAE achieved the highest log-likelihood. Although  $\mathcal{WB}$ -VAE and  $\mathcal{MWB}$ -VAE did not surpass PoE-VAE in log-likelihood, their log-likelihoods were on par with MoPoE-VAE.

**CelebA results.** As shown in Table 2, the proposed  $\mathcal{WB}$ -VAE outperformed PoE-VAE as well as competed favorably and even better than MoE-VAE in both latent representation and generation on the challenging bimodal CelebA dataset. Likewise,  $\mathcal{MWB}$ -VAE outperformed MoPoE-VAE in most scenarios, with the exception of latent representation classification when using image as the input modality. As consistent with the trends observed in the previous two datasets, the latent representation classification accuracy of  $\mathcal{WB}$ -VAE increased as more modalities were present, similar to PoE-VAE. In contrast, the classification accuracy of MoE-VAE decreased when more modalities were given. Remarkably, both  $\mathcal{WB}$ -VAE and  $\mathcal{MWB}$ -VAE achieved good performance for the most challenging image-to-text generation task, outperforming the second-best method by 11.5% and 30.8%, respectively.  $\mathcal{MWB}$ -VAE also achieved good performance in text-to-image conditional generation (see Fig. 4), where  $\mathcal{MWB}$ -VAE learned good representations of different attributes well (*e.g.*, "smiling," "hairstyles," etc).

## Conclusion

In this work, we introduced a barycentric perspective on previous multimodal VAEs, offering a theoretical and unified formulation. This approach allows for explorations of various aggregation functions in the regime of multimodal VAEs. Leveraging this barycentric formulation, we proposed a  $\mathcal{WB}$ -VAE, which uses the Wasserstein barycenter as an aggregation function that better preserves the geometry of unimodal distributions. Experimental results showed the effectiveness of the proposed  $\mathcal{WB}$ -VAE when compared to other state-of-the-art multimodal VAEs. We hope our new perspective will stimulate the exploration of other aggregation functions for multimodal VAEs in future work.

## Acknowledgments

This work was partially supported by NIH grant R01-AG067103. Computations were performed using the resources of the Washington University Research Computing and Informatics Facility, which were partially funded by NIH grants S10OD025200, 1S10RR022984-01A1 and 1S10OD018091-01.

## References

- Agueh, M.; and Carlier, G. 2011. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924.
- Ambrosio, L.; Gigli, N.; and Savaré, G. 2008. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.
- Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. [arXiv:1701.07875](https://arxiv.org/abs/1701.07875).
- Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.
- Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.
- Givens, C. R.; and Shortt, R. M. 1984. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2): 231–240.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Grove, K.; and Karcher, H. 1973. How to conjugate  $C^1$ -close group actions. *Mathematische Zeitschrift*, 132: 11–20.
- Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.
- Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C. P.; Bosnjak, M.; Shanahan, M.; Botvinick, M.; Hassabis, D.; and Lerchner, A. 2017. Scan: Learning hierarchical compositional visual concepts. [arXiv preprint arXiv:1707.03389](https://arxiv.org/abs/1707.03389).
- Hirt, M.; Campolo, D.; Leong, V.; and Ortega, J.-P. 2024. Learning multi-modal generative models with permutation-invariant encoders and tighter variational objectives. *Transactions on Machine Learning Research*.
- Kantorovich, L. V. 1942. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, 199–201.
- Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. [arXiv preprint arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
- Knott, M.; and Smith, C. S. 1984. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43: 39–49.
- Knott, M.; and Smith, C. S. 1994. On a generalization of cyclic monotonicity and distances among random vectors. *Linear algebra and its applications*, 199: 363–371.
- 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.
- LeCun, Y.; and Cortes, C. 2010. MNIST handwritten digit database.
- Lin, Y.-B.; Sung, Y.-L.; Lei, J.; Bansal, M.; and Bertasius, G. 2023. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2299–2309.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.
- Minka, T.; et al. 2005. Divergence measures and message passing. Technical report, Technical report, Microsoft Research.
- Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. [arXiv preprint arXiv:1802.05957](https://arxiv.org/abs/1802.05957).
- Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 666–704.
- Murphy, K. P. 2012. *Machine learning: a probabilistic perspective*. MIT press.
- 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*, volume 2011, 4. Granada.
- Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.
- Palumbo, E.; Daunhawer, I.; and Vogt, J. E. 2023. MM-VAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview.
- Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.
- Pham, H.; Liang, P. P.; Manzi, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6892–6899.
- 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.
- 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.

Sutter, T.; Daunhawer, I.; and Vogt, J. 2020. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33: 6100–6110.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized multimodal ELBO. *arXiv preprint arXiv:2105.02470*.

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. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Turner, R.; and Sahani, M. 2011. Two problems with variational expectation maximisation for time-series models. Cambridge University Press.

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

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

Yuan, S.; Cui, J.; Li, H.; and Han, T. 2024. Learning Multimodal Latent Generative Models with Energy-Based Prior. *arXiv preprint arXiv:2409.19862*.