Multimodal Gaussian Mixture Variational Autoencoder with Consistency Regularizations

Yarui Chen[1], Lehan Hong[1], Jianlin Shao[1], Jianning Yang[2], Tingting Zhao[1*], Yun Liao[1], Yancui Shi[1*]

[1]Tianjin University of Science and Technology
[2]Xi'an University of Posts and Telecommunications
{yrchen, tingting, yliao, syc}@tust.edu.cn, {hlh, Sjl_}@mail.tust.edu.cn, yangjn@stu.xupt.edu.cn

Abstract

Variational autoencoder (VAE)-based frameworks possess a natural advantage in modeling the shared and private information inherent in multimodal data. However, current models focus on improving the quality of shared representations from the reconstruction perspective, lacking explicit mechanisms to model their underlying semantic structure. In this paper, we propose the multimodal Gaussian mixture variational autoencoder with consistency regularizations, which introduces a Gaussian mixture prior over the shared latent space to enhance its semantic structure and encourage the formation of cluster-aware latent representations. To address the cross-modal inconsistency problem under missing modality conditions, we propose a cluster-guided regularization strategy that enforces the cross-modal consistency using the pseudo-category labels from unsupervised clustering. Additionally, we design a self-supervised contrastive regularization strategy to align semantically similar representations across modalities. Extensive experiments on MNIST-SVHN and MNIST-CDCB datasets demonstrate that our method significantly outperforms prior state-of-the-art models in generation, classification, and retrieval tasks.

Introduction

With the increasing availability of multimodal data, learning effective representations across heterogeneous modalities has become a fundamental challenge in machine learning (Bengio, Courville, and Vincent 2013; Baltrušaitis, Ahuja, and Morency 2018; Zhan et al. 2023). Both Variational Autoencoders (VAEs) (Kingma and Welling 2013; Wu and Goodman 2019; Korthals et al. 2019; Bachmann et al. 2022; Bouchacourt, Tomioka, and Nowozin 2018; Daunhawer et al. 2021) and Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Liu et al. 2024; Zhu et al. 2024; Meng et al. 2025; Zhang et al. 2021; Zhou et al. 2023) have been widely adopted in multimodal generative modeling. Compared to GANs, VAEs offer a principled probabilistic framework that is particularly suitable for learning structured and disentangled representations. Besides, VAEs are more stable and interpretable, making them well-suited for modeling the joint distribution of heterogeneous modalities in a unified latent space.

A key goal of multimodal representation learning is to construct a shared latent space that captures cross-modal semantic consistency while preserving modality-specific characteristics (Suzuki and Matsuo 2022). Early multimodal generative models, such as JMVAE (Suzuki, Nakayama, and Matsuo 2016), learn a shared latent space by jointly encoding all modalities. However, this approach requires all modalities to be present during both training and inference, making it unsuitable for real-world scenarios with missing data. To improve flexibility, MVAE (Wu and Goodman 2018) introduces a Product-of-Experts (PoE) (Hinton 1999) inference framework that allows separate encoding of each modality and robust fusion under partial observations. MMVAE (Shi et al. 2019) further adopts a Mixture-of-Experts (MoE) strategy to adaptively weight each modality's contribution. MoPoE-VAE (Sutter, Daunhawer, and Vogt 2021) introduces a generalized ELBO formulation that unifies the PoE and MoE paradigms by modeling the joint posterior as a Mixture-of-Products-of-Experts. Despite their flexibility in handling missing modalities, these models largely neglect the presence of both shared and modality-specific information.

Subsequent methods have explored the disentanglement of shared and private representations in multimodal data (Daunhawer et al. 2020; Hwang et al. 2020). DMVAE (Lee and Pavlovic 2021) explicitly separates the latent space by minimizing the mutual information between shared and private components, and aligns shared representations across modalities using the InfoNCE loss (van den Oord, Li, and Vinyals 2018). SSDMM-VAE (Mondal et al. 2023) introduces a semi-supervised framework that disentangles shared and private representations by modeling latent factors with both discrete and continuous components. The model enhances shared representational expressiveness through a combination of PoE inference and statistical ensemble learning. MMVAE+ (Palumbo, Daunhawer, and Vogt 2023) decouples the latent space into a shared subspace for encoding cross-modal semantic information and modality-specific subspaces for capturing details. And it uses auxiliary priors instead of modality-specific encodings, which forces decoders to rely on the shared latent subspace.

*Corresponding author.

Despite recent progress, existing multimodal generative models focus on improving the quality of shared representations from a data reconstruction perspective, often overlooking their underlying semantic structure and cross-modal consistency. In this work, we explicitly model the semantic structure of shared representations by introducing a Gaussian Mixture prior over the joint shared latent vector and encouraging the learned shared space to be semantically meaningful and cluster-aware. Furthermore, we propose two regularization strategies to enforce the consistency of shared representations across modalities.

Specifically, we propose the MGMVAE (Multimodal Gaussian Mixture Variational Autoencoder with consistency regularizations), which employs a variational framework to extract shared and private information from multimodal data. The model imposes a Gaussian mixture prior over the shared latent space to enhance its clustering structure. To further improve cross-modal consistency, we introduce a cluster-guided regularization that externally aligns shared representations across modalities via the joint latent vector. Additionally, we design a self-supervised contrastive regularization that promotes semantic alignment by encouraging similarity between positive modality pairs while discriminating against negative ones.

The main contributions of this work are summarized as follows:

- We propose the MGMVAE, a novel multimodal variational autoencoder that explicitly disentangles shared and private latent vectors. The model incorporates a Gaussian Mixture prior to induce clustering structure in the joint shared latent vector.
- We introduce two complementary regularization strategies to enhance cross-modal consistency: a cluster-guided regularization that leverages pseudo-labels from the joint latent space to guide alignment, and a self-supervised contrastive regularization that enforces semantic similarity across modalities.
- Extensive experiments on MNIST-SVHN and MINST-CDCB datasets demonstrate that the MGMVAE consistently outperforms state-of-the-art baselines in unsupervised multimodal generation, classification and retrieval tasks.

Related Works

Multimodal Variational Autoencoders

Multimodal VAEs have become a key framework for learning joint latent representations from heterogeneous modalities. Various extensions of multimodal VAEs based on the disentanglement of shared and modality-specific information have been proposed to improve shared representation learning under challenging conditions (Senellart and Allassonnière 2025; Yuan, Lipizzi, and Han 2024).

Some methods enhance the flexibility of shared representation learning by adopting adaptive alignment mechanisms (Hwang et al. 2020; Sutter et al. 2024). MMVM-VAE (Sutter et al. 2024) employs a PoE prior to softly align modality-specific latent representations toward a shared aggregate posterior. It encourages cross-modal similarity via minimizing Jensen-Shannon divergence between unimodal posteriors. CMMD (Mancisidor et al. 2024) employs marginal Maximum Mean Discrepancy (MMD) regularization to align the average posterior distribution with the conditional prior. This mechanism ensures coherent shared representations across varying modality combinations.

Numerous studies achieve consistency by aggregating unimodal posteriors through flexible optimization of distances and expressive priors (Qiu et al. 2025; Oshima et al. 2024; Yuan et al. 2024). Unlike traditional KL divergence-based aggregation, $\mathcal{MWB}$-VAE (Qiu et al. 2025) uses the Wasserstein barycenter to aggregate distributions, preserving their geometric structure. By minimizing the 2-Wasserstein distance, it better preserves the latent geometry of unimodal distributions and captures both modality-specific and shared information. Yuan et al. (Yuan et al. 2024) replace rigid Gaussian priors with an energy-based model for the shared subspace. It leverages variational inference combined with MCMC sampling to capture complex cross-modal dependencies and enhance the consistency of the learned shared representations.

Variational Deep Generative Clustering Models

Variational deep generative clustering models aim to structure the latent space into distinct clusters that reflect meaningful data groupings. This not only enhances the interpretability of latent representations but also enables accurate clustering and high-quality data generation.

Early work like VaDE (Jiang et al. 2016) extends the VAE framework by incorporating a Gaussian mixture model prior in the latent space, enabling simultaneous unsupervised clustering and generative modeling. Recent extensions apply latent clustering to multimodal settings (Wang et al. 2021; Palumbo et al. 2023; Hirt et al. 2023). DVIMC (Xu et al. 2024) employs a PoE framework combined with a coherence objective to aggregate information from incomplete views. CMVAE (Palumbo et al. 2024) incorporates a mixture prior into the latent space to enhance clustering and generative quality. It also introduces mechanisms for automatic cluster number selection and leverages diffusion probabilistic models to improve generation fidelity.

Clustering methods have demonstrated substantial advantages in both unimodal and multimodal representation learning, such as capturing underlying semantics, enhancing interpretability, and improving generalization.

Methods

We propose the MGMVAE, which proposes a Gaussian mixture prior to model the shared semantic structures, and designs consistency regularizations to align shared representations across modalities and enhance the coherence of the latent space. The generative and inference processes are illustrated in Figure 1. As shown in Figure 1(a), we suppose the observations $\boldsymbol{x}$ and $\boldsymbol{y}$ are generated from $\boldsymbol{z}_s$ with a Gaussian mixture prior c, together with their respective private latent vectors $\boldsymbol{h}_x$ and $\boldsymbol{h}_y$. During the inference process illustrated in Figure 1(b), each modality is encoded into both

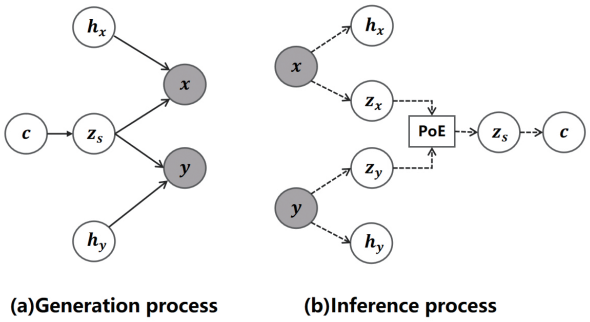

Figure 1: The schematic diagram of model generation and inference process.

shared and private latent vectors. The shared latent vectors $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ are combined using the PoE approach to form $\boldsymbol{z}_s$.

Multimodal Architecture

We take a bimodal data $(\boldsymbol{x}, \boldsymbol{y})$ as input. The joint shared latent $\boldsymbol{z}_s$ denotes the joint shared latent vector and $\boldsymbol{h}_x$, $\boldsymbol{h}_y$ represent the private latent vectors for $\boldsymbol{x}$, $\boldsymbol{y}$. To enhance the clustering accuracy of the joint shared latent space, we place a Gaussian mixture prior over $\boldsymbol{z}_s$, with K components and a discrete cluster assignment variable $\boldsymbol{c} \in \{1, \ldots, K\}$. The joint probability distribution of the MGMVAE is defined as:

$$\begin{aligned} &p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c}) = \\ &p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{h}_x)\, p_{\boldsymbol{\theta}_2}(\boldsymbol{y}|\boldsymbol{z}_s, \boldsymbol{h}_y)\, p(\boldsymbol{z}_s|\boldsymbol{c})\, p(\boldsymbol{c})\, p(\boldsymbol{h}_x)\, p(\boldsymbol{h}_y) \end{aligned} \tag{1}$$

where $p(\boldsymbol{c}) = \text{Cat}(\boldsymbol{\pi})$ is a categorical prior over cluster assignments, and $p(\boldsymbol{z}_s|\boldsymbol{c} = k) = \mathcal{N}(\boldsymbol{z}_s|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ defines the Gaussian components. Private latent vectors follow standard Gaussian priors. $p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{h}_x)$ and $p_{\boldsymbol{\theta}_2}(\boldsymbol{y}|\boldsymbol{z}_s, \boldsymbol{h}_y)$ denote the conditional likelihoods of $\boldsymbol{x}$ and $\boldsymbol{y}$, parameterized by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

The marginal likelihood of a single observation pair $(\boldsymbol{x}, \boldsymbol{y})$ under the generative model is given by:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{\boldsymbol{c}=1}^{K} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c})\, d\boldsymbol{z}_s\, d\boldsymbol{h}_x\, d\boldsymbol{h}_y \\ &= \sum_{\boldsymbol{c}=1}^{K} \int p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{h}_x)\, p_{\boldsymbol{\theta}_2}(\boldsymbol{y}|\boldsymbol{z}_s, \boldsymbol{h}_y) \\ &\qquad p(\boldsymbol{z}_s|\boldsymbol{c})\, p(\boldsymbol{c})\, p(\boldsymbol{h}_x)\, p(\boldsymbol{h}_y)\, d\boldsymbol{z}_s\, d\boldsymbol{h}_x\, d\boldsymbol{h}_y \end{aligned} \tag{2}$$

Since this marginal likelihood is intractable in general, we introduce a variational distribution $q(\boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c}|\boldsymbol{x}, \boldsymbol{y})$ to approximate the true posterior. By applying variational inference, we obtain the evidence lower bound (ELBO) of the log-likelihood:

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) &\geq E_q \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c})}{q(\boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c}|\boldsymbol{x}, \boldsymbol{y})} \right] \\ &= \mathcal{L}_{\text{ELBO}}(\boldsymbol{x}, \boldsymbol{y}) \end{aligned} \tag{3}$$

The variational posterior is defined as:

$$\begin{aligned} q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{h}_x, \boldsymbol{h}_y, \boldsymbol{c}|\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{z}_s|\boldsymbol{x}, \boldsymbol{y})\, q_{\boldsymbol{\phi}_1}(\boldsymbol{h}_x|\boldsymbol{x}) \\ \cdot q_{\boldsymbol{\phi}'_1}(\boldsymbol{h}_y|\boldsymbol{y})\, q(\boldsymbol{c}|\boldsymbol{z}_s) \end{aligned} \tag{4}$$

where $q(\boldsymbol{z}_s|\boldsymbol{x}, \boldsymbol{y})$ denotes the approximate posterior of the joint shared latent vector $\boldsymbol{z}_s$, $q_{\boldsymbol{\phi}_1}(\boldsymbol{h}_x|\boldsymbol{x})$ and $q_{\boldsymbol{\phi}'_1}(\boldsymbol{h}_y|\boldsymbol{y})$ denote the approximate posteriors of the modality private latent vectors $\boldsymbol{h}_x$ and $\boldsymbol{h}_y$, and $q(\boldsymbol{c}|\boldsymbol{z}_s)$ denotes the posterior distribution over the discrete cluster assignment, which is typically computed using Bayes' rule under the GMM prior.

To support partial modality inference, we use the PoE-based fusion:

$$q(\boldsymbol{z}_s|\boldsymbol{x}, \boldsymbol{y}) \propto p(\boldsymbol{z}_s)\, q_{\boldsymbol{\phi}_2}(\boldsymbol{z}_x|\boldsymbol{x})\, q_{\boldsymbol{\phi}'_2}(\boldsymbol{z}_y|\boldsymbol{y}) \tag{5}$$

where $p(\boldsymbol{z}_s)$ denotes the prior distribution over $\boldsymbol{z}_s$, and $q_{\boldsymbol{\phi}_2}(\boldsymbol{z}_x|\boldsymbol{x})$ and $q_{\boldsymbol{\phi}'_2}(\boldsymbol{z}_y|\boldsymbol{y})$ are the modality approximate posteriors of the shared latent vectors $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$, inferred from modalities $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. When a modality is missing, the corresponding posterior term is omitted, enabling robust inference under partial observations.

The resulting ELBO objective becomes:

$$\begin{aligned} &\mathcal{L}_{\text{ELBO}} = \\ &E_{q(\boldsymbol{z}_s, \boldsymbol{h}_x)} \left[\log p(\boldsymbol{x}|\boldsymbol{z}_s, \boldsymbol{h}_x)\right] + E_{q(\boldsymbol{z}_s, \boldsymbol{h}_y)} \left[\log p(\boldsymbol{y}|\boldsymbol{z}_s, \boldsymbol{h}_y)\right] \\ &- \lambda_1 \Bigg(E_{q(\boldsymbol{z}_s)}[D_{\text{KL}}\big(q(\boldsymbol{c}|\boldsymbol{z}_s)\|p(\boldsymbol{c})\big)] \\ &\quad + \sum_{\boldsymbol{c}} q(\boldsymbol{c}|\boldsymbol{z}_s) D_{\text{KL}}\big(q(\boldsymbol{z}_s|\boldsymbol{x}, \boldsymbol{y})\|p(\boldsymbol{z}_s|\boldsymbol{c})\big) \Bigg) \\ &- \lambda_2 \Big[D_{\text{KL}}\big(q(\boldsymbol{h}_x|\boldsymbol{x})\|p(\boldsymbol{h}_x)\big) + D_{\text{KL}}\big(q(\boldsymbol{h}_y|\boldsymbol{y})\|p(\boldsymbol{h}_y)\big) \Big] \\ &- \lambda_3 \Big[D_{\text{KL}}\big(q(\boldsymbol{z}_x|\boldsymbol{x})\|p(\boldsymbol{z}_x)\big) + D_{\text{KL}}\big(q(\boldsymbol{z}_y|\boldsymbol{y})\|p(\boldsymbol{z}_y)\big) \Big] \end{aligned} \tag{6}$$

where λ_1, λ_2, and λ_3 are balancing hyperparameters.

Cluster-guided Regularization

To enhance the consistency of cross-modal semantics, we propose a cluster-guided regularization term guided by pseudo-labels obtained from the underlying clustering structure. This regularization encourages samples assigned to the same cluster to be pulled toward a corresponding learnable center in the shared latent space. This approach promotes consistent latent alignment and reinforces inter-class separability across modalities with varying data quality.

A cluster assignment distribution $q(c|\boldsymbol{z}_s)$ is computed using a variational Gaussian mixture model (GMM), and each sample is assigned a pseudo-label corresponding to the most probable cluster:

$$\hat{c}^{(i)} = \arg\max_k q(c = k \mid \boldsymbol{z}_s^{(i)}) \tag{7}$$

To reduce the impact of uncertain or noisy cluster assignments during training, we adopt a confidence-aware sample selection strategy. The confidence score for each sample is defined as the highest posterior probability of its assigned cluster:

$$\text{conf}_i = \max_k q(c = k \mid \boldsymbol{z}_s^{(i)}) \tag{8}$$

Only samples with confidence exceeding a predefined threshold $\tau \in [0, 1]$ are retained to guide clustering-based

alignment. Let $H = \{i \mid \text{conf}_i > \tau\}$ denote the index set of high-confidence samples.

For those selected samples, we apply a center loss to the shared latent vectors $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$, encouraging them to be close to the respective cluster centers:

$$\mathcal{L}_{\text{center}} = \frac{1}{|H|} \sum_{i \in H} \left(\left\| \boldsymbol{z}_x^{(i)} - \boldsymbol{c}_{\hat{c}^{(i)}} \right\|_2^2 + \left\| \boldsymbol{z}_y^{(i)} - \boldsymbol{c}_{\hat{c}^{(i)}} \right\|_2^2 \right) \tag{9}$$

where $\boldsymbol{c}_{\hat{c}^{(i)}}$ is the learnable center associated with cluster $\hat{c}^{(i)}$, and $\{\boldsymbol{c}_k\}_{k=1}^K$ denotes the set of cluster centers jointly optimized during training to minimize intra-class variance.

By leveraging the joint shared latent vector $\boldsymbol{z}_s$ to generate pseudo-labels and selectively applying center loss to high-confidence samples, our approach enforces a consistent clustering structure across modalities. This design enhances semantic coherence and structural alignment in the learned representations, particularly under unsupervised multimodal settings.

Self-supervised Contrastive Regularization

To ensure alignment of shared latent representations across modalities, we design a self-supervised contrastive regularization that operates on paired training data and incorporates selective negative sampling to improve the discriminative quality of the learned representations. By pulling together shared latent vectors from modality pairs that correspond to the same semantic concept and pushing apart those from different concepts, the model learns a semantically consistent latent space.

Formally, given a triplet consisting of a positive pair (shared latent vectors from two modalities corresponding to the same instance) and negative samples (shared latent vectors from different instances), the alignment loss enforces that the distance between the positive pair is smaller than that between the positive and negative pairs, i.e.,

$$d\big(\boldsymbol{\mu}_1(\boldsymbol{x}),\ \boldsymbol{\mu}_2(\boldsymbol{y})\big) + \alpha < d\big(\boldsymbol{\mu}_1(\boldsymbol{x}),\ \boldsymbol{\mu}_2(\boldsymbol{y}')\big) \tag{10}$$

where $d(\cdot, \cdot)$ denotes a distance metric—specifically, the Euclidean distance in our case. Here, $\boldsymbol{\mu}_1(\boldsymbol{x})$ and $\boldsymbol{\mu}_2(\boldsymbol{y})$ represent the means of the shared latent vectors encoded from modality x and modality y, respectively, and α is a margin hyperparameter controlling the separation between positive and negative pairs.

To optimize this objective, we employ the standard triplet loss:

$$\mathcal{L}_{\text{triplet}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}') = \max\Big\{ \|\boldsymbol{\mu}_1(\boldsymbol{x}) - \boldsymbol{\mu}_2(\boldsymbol{y})\|_2 - \frac{1}{K} \sum_{i=1}^{K} \big\|\boldsymbol{\mu}_1(\boldsymbol{x}) - \boldsymbol{\mu}_2(\boldsymbol{y}'^i)\big\|_2 + \alpha,\ 0 \Big\} \tag{11}$$

where K denotes the number of negative samples per training pair. This loss encourages shared latent vectors of positive pairs to be close, effectively aligning the shared latent space across modalities, while pushing apart negative pairs to facilitate the disentanglement of shared and private latent vectors.

Accordingly, the multimodal shared latent alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{triplet}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}') + \mathcal{L}_{\text{triplet}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{y}') \tag{12}$$

where $\boldsymbol{x}'$ and $\boldsymbol{y}'$ denote negative samples for modalities x and y, respectively.

During training, the selection of negative samples for each training pair is critical. we adopt a two-stage self-supervised negative sampling strategy. Initially, negative samples are randomly chosen to allow the model to learn coarse representations despite noisy assignments. As training progresses, we leverage the partially trained model to dynamically filter negative samples based on a semi-hard negative mining criterion. Specifically, we retain only those negatives whose distance to the anchor is greater than that of the positive pair, ensuring they are informative yet not trivially separable. This progressive refinement improves both training stability and alignment quality in the shared latent space.

Overall Objective Loss

Combining the above components with appropriate weighting coefficients, the overall objective of MGMVAE integrates multiple complementary loss terms to jointly optimize representation learning, clustering consistency, and cross-modal alignment. Formally, the final training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \beta \mathcal{L}_{\text{center}} + \gamma \mathcal{L}_{\text{align}} \tag{13}$$

where the hyperparameters β and γ balance the contributions of the respective loss terms.

Figure 2 illustrates the overall architecture of our proposed MGMVAE model. The framework is composed of two encoders and decoders, which are used to extract both shared and private latent representations from each modality. The PoE mechanism is employed to fuse unimodal posteriors into a joint posterior over the shared latent space. To enhance the consistency and alignment of shared representations, we introduce two complementary regularization strategies: a cluster-guided regularization and a self-supervised contrastive regularization.

Experiments

In this section, we conduct comprehensive experiments on MNIST-SVHN and MNIST-CDCB datasets to evaluate the effectiveness of our MGMVAE model. Firstly, we perform cross-modal generation tasks to assess the model's ability to capture modality-invariant semantic representations. Secondly, we design modality translation experiments to evaluate the model's capability to accurately transfer semantic content across different modalities. Moreover, we assess the quality of the learned shared latent representations on downstream tasks, including classification and retrieval, which demonstrates the model's robustness and discriminative power in practical multimodal applications.

Datasets

We evaluate our model on two widely used multimodal generative benchmarks: MNIST-SVHN (Shi et al. 2019)

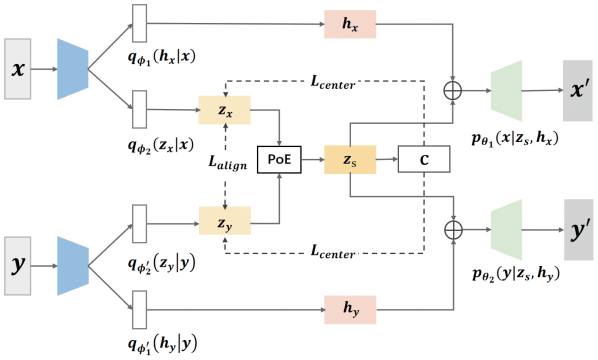

Figure 2: The framework of of the MGMVAE model. The model employs a shared decoder to disentangle private and shared information for each modality. Here, $\mathcal{L}_{\text{center}}$ denotes the cluster-guided regularization, while $\mathcal{L}_{\text{align}}$ represents the self-supervised contrastive regularization.

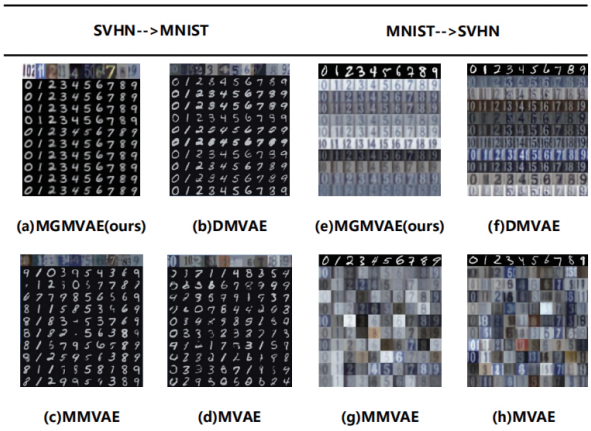

Figure 3: Cross-modal generation results on the MNIST-SVHN dataset. Given an input from one modality, the model reconstructs the corresponding sample in the other modality.

and MNIST-CDCB (Gonzalez-Garcia, Weijer, and Bengio 2018), both designed to assess cross-modal representation learning and generation under paired settings.

MNIST-SVHN combines handwritten digits from MNIST (LeCun et al. 1998) and street-view digits from SVHN (Netzer et al. 2011), forming a challenging scenario due to significant variations in style, resolution, and background.

MNIST-CDCB is a synthetic dataset that pairs grayscale MNIST digits with transformed versions featuring colorization, distortion, and complex backgrounds.

Together, these datasets cover both real-world and synthetic modality gaps, enabling comprehensive evaluation of multimodal disentanglement, alignment, and generation capabilities.

Experiment Settings

For MNIST-SVHN, the model is trained for 20 epochs using a batch size of 128. For MNIST-CDCB, we train for 50 epochs with a batch size of 64. In both cases, we set the latent dimensionality to 20 and use the Adam (Kingma 2014) optimizer with a learning rate of 0.0001. In addition, the threshold τ is set to 0.8.

The loss-related hyperparameters are empirically chosen as follows: center loss weight $\beta = 1$, alignment loss weight $\gamma = 1$, and $\lambda_1 = 0.01$, $\lambda_2 = 0.01$. For MNIST-SVHN, we set $\lambda_3 = 0.0025$, and for MNIST-CDCB, $\lambda_3 = 0.001$. The number of negative samples per triplet is fixed at $K = 1$, and the margin hyperparameter is set to 1.7 for MNIST-SVHN and 2.0 for MNIST-CDCB.

Cross-modal Generation

Our model supports cross-modal generation, where data from one modality can be generated conditioned solely on the observation from the other modality. This capability highlights the model's effectiveness in capturing shared representations and performing robust inference under partial modality scenarios. By leveraging modality-invariant semantics, the model can reconstruct or complete missing modality data using the available input.

MNIST-SVHN We evaluate this functionality on the MNIST-SVHN dataset by comparing our model with several representative multimodal generative baselines, including MVAE (Wu and Goodman 2018), MMVAE (Shi et al. 2019), DMVAE (Lee and Pavlovic 2021), MMVAE+ (Palumbo, Daunhawer, and Vogt 2023), CMMD (Mancisidor et al. 2024), and $\mathcal{MWB}$-VAE (Qiu et al. 2025). Evaluations are conducted in both generation directions.

The quantitative results of cross-modal generation accuracy are reported in Table 1, where accuracy is evaluated using a pretrained digit classifier to verify class identity preservation. Our model consistently outperforms all baselines in both directions. In particular, it shows a notable improvement in the more challenging SVHN→MNIST direction, where clean digit generation must overcome background clutter and visual noise in the SVHN inputs. This result underscores the model's robustness to modality imbalance and its effectiveness in learning modality-invariant, class-consistent representations. The qualitative results are shown in Figure 3. The first row shows the observed input modality, and subsequent rows depict generated samples from each model under missing-modality conditions. Our model produces visually clearer, semantically accurate digits with sharper contours and fewer artifacts, confirming its effectiveness in generating structurally consistent and high-fidelity cross-modal samples.

MNIST-CDCB To further evaluate the model's ability to disentangle shared and private representations, we perform cross-modal generation experiments on the MNIST-CDCB dataset. This dataset includes two complementary modalities—MNIST-CD and MNIST-CB—containing digit images with distinct background styles and color distributions.

As shown in Figure 4, the left panel illustrates the generation results from MNIST-CD to MNIST-CB. Despite the

Model	MNIST → SVHN	SVHN → MNIST
MVAE(Wu and Goodman 2018)	31.8	57.1
MMVAE+(Palumbo, Daunhawer, and Vogt 2023)	69.0	62.0
MMVAE(Shi et al. 2019)	80.0	70.0
CMMD(Mancisidor et al. 2024)	75.0	66.0
$\mathcal{MWB}$-VAE(Qiu et al. 2025)	82.0	36.0
DMVAE(Lee and Pavlovic 2021)	<u>91.6</u>	<u>76.4</u>
MGMVAE (Ours)	**95.4**	**86.7**

Table 1: Cross-modal generation accuracy (%) on the MNIST-SVHN dataset. **Bold** indicates the best performance, and <u>underline</u> denotes the second best.

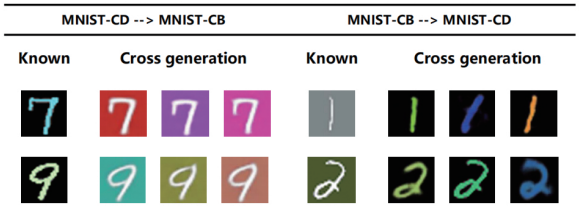

Figure 4: Cross-modal generation results on the MNIST-CDCB dataset.

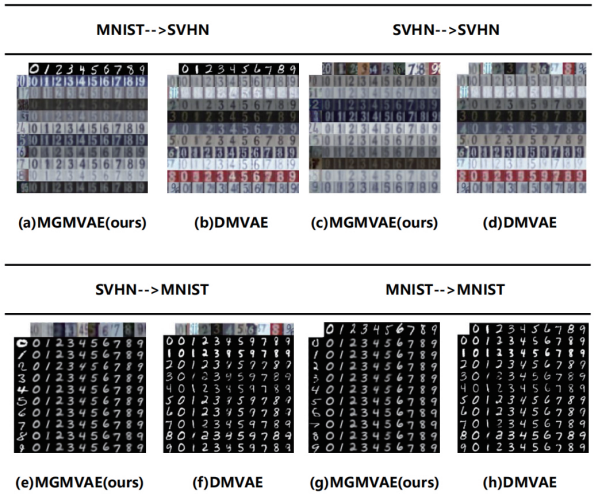

Figure 5: Translation generation results across modalities s on the MNIST-SVHN dataset. The model takes an input from one modality and translates it into the corresponding sample in the other modality.

complex and randomly colored backgrounds in the target modality, the model consistently reconstructs digits in white, preserving the semantic and structural content of the input. The right panel shows the reverse direction—from MNIST-CB to MNIST-CD—where digit shapes are well preserved, and visual characteristics are accurately adapted to match the grayscale appearance of the target modality.

These results on the MNIST-SVHN and MNIST-CDCB datasets demonstrate that the proposed model effectively captures the shared semantic features while disentangling them from the modality-specific factors. The generated samples maintain input semantics and adapt to modality-specific attributes that indicate strong cross-modal generalization and robustness under missing-modality conditions.

Translation Generation

The objective of translation generation is to synthesize samples in the style of a target modality using semantic content from a source modality.

MNIST-SVHN We conduct experiments on the MNIST-SVHN dataset and compare our method with the strongest baseline, DMVA (Mondal et al. 2023). As shown in Figure 5, our model consistently generates more accurate and visually coherent samples across different translation directions. Compared to DMVAE, our model achieves better disentanglement between shared and private representations, which allows it to more accurately transfer style characteristics from the reference modality while maintaining the correct digit identity from the source input.

MNIST-CDCB In addition to MNIST-SVHN, we also evaluate our model on the MNIST-CDCB dataset. As illustrated in Figure 6, our model successfully performs bidirectional translation between the MNIST-CD and MNIST-CB modalities, preserving digit semantics while adapting to distinct visual styles.

These results on the MNIST-SVHN and MNIST-CDCB datasets highlight the superiority of our framework in learning robust and transferable multimodal representations, which lead to improved generation quality and stronger modality adaptability compared to prior methods.

Downstream Tasks

Multimodal Data Classification Our model supports multimodal data classification by utilizing shared latent vectors extracted from paired training samples. For each input pair, the encoders produce shared latent vectors, which are then fused into a joint shared latent vector through a PoE mechanism. A linear classifier is trained on the joint shared latent vector alongside its corresponding class labels. During inference, we extract the shared latent vectors from the test data and feed them into the linear classifier to perform label prediction.

We evaluate our approach on the MNIST-SVHN dataset and compare classification accuracy against representative

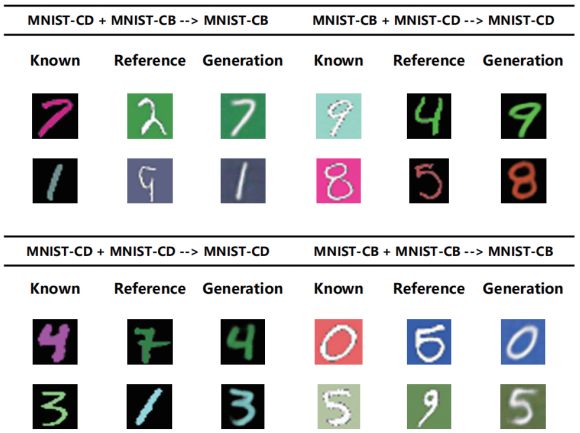

Figure 6: Translation generation results across modalities s on the MNIST-CDCB dataset.

Model	MNIST	SVHN
MVAE(Wu and Goodman 2018)	79.80	65.10
MMVAE(Shi et al. 2019)	91.30	68.00
DMVAE(Lee and Pavlovic 2021)	95.00	79.90
$\mathcal{MWB}$-VAE(Qiu et al. 2025)	97.00	83.00
(Oshima et al. 2024)	98.15	78.50
MGMVAE (Ours)	**98.62**	**86.79**

Table 2: Classification accuracy (%) on MNIST and SVHN modalities.

multimodal generative models, including MVAE (Wu and Goodman 2018), MMVAE (Shi et al. 2019), DMVAE (Lee and Pavlovic 2021), $\mathcal{MWB}$-VAE (Qiu et al. 2025) and (Oshima et al. 2024). The classifier is a single-layer neural network with an input dimension of 20 and an output dimension of 10.

As summarized in Table 2, our method outperforms all baselines across both modalities. This indicates that the learned shared latent space captures consistent and semantically meaningful features, which enables robust generalization to unseen test samples.

Cross-modal Data Retrieval Our model is also applicable to cross-modal retrieval that serves as a downstream task to assess the semantic quality of the learned shared representations. We evaluate this capability on the MNIST-CDCB dataset. In this task, all samples from the retrieval database are encoded into their corresponding shared latent vectors and grouped by modality. At test time, a query sample from one modality is encoded into its shared latent vector, and pairwise distances are computed between the query and all target-modality latent vectors. Retrieval results are ranked by ascending distance, with top-ranked samples considered most semantically similar.

Retrieval accuracy is evaluated based on whether the top-1 retrieved sample shares the same digit label as the query. As shown in Table 3, our model outperforms CdDN (Lee et al. 2018) and IIAE (Hwang et al. 2020). This confirms the effectiveness of our method in aligning multimodal representations for reliable semantic retrieval under modality shifts.

Model	CD $\rightarrow$ CB	CB $\rightarrow$ CD
CdDN(Lee et al. 2018)	99.6	99.6
IIAE(Hwang et al. 2020)	99.7	99.7
MGMVAE (Ours)	**99.9**	**99.9**

Table 3: Cross-modal retrieval accuracy (%) on the MNIST-CDCB dataset.

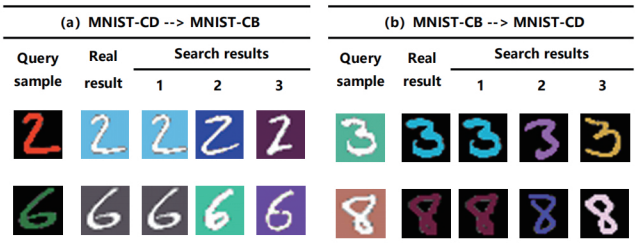

Figure 7: Experimental results of cross-modal data retrieval: (a) MNIST-CD to MNIST-CB, (b) MNIST-CB to MNIST-CD.

Figure 7(a) presents cross-modal retrieval results from MNIST-CD to MNIST-CB. In each example, the first column corresponds to the query image, the second column shows the ground-truth counterpart, and the third to fifth columns display the top-1 to top-3 retrieved results. The top-1 retrievals consistently match the ground truth in terms of digit identity, while the top-2 and top-3 results also exhibit strong semantic consistency. Comparable trends are observed in Figure 7(b), which illustrates retrieval in the reverse direction. These results demonstrate that our model learns modality-invariant latent representations capable of robust cross-modal semantic matching.

Conclusion

In this work, we propose the MGMVAE, a multimodal variational autoencoder that jointly models shared and private representations across heterogeneous modalities. To enhance the semantic structure of the shared latent space, the model incorporates a Gaussian mixture prior over the joint shared representation. By integrating cluster-guided and self-supervised contrastive consistency regularizations, MGMVAE effectively aligns cross-modal representations. Extensive experiments conducted on the MNIST-SVHN and MNIST-CDCB datasets demonstrate the superior performance of the MGMVAE across multiple tasks, including generation, classification, and retrieval. In particular, under the presence of imbalanced data quality—where the SVHN dataset contains noisier backgrounds—MGMVAE effectively extracts modality-invariant semantic features. This robustness to noisy inputs leads to significantly improved accuracy in the SVHN$\rightarrow$MNIST cross-modal generation task.

Acknowledgments

This work was supported by Tianjin Science and Technology Plan Project [25JCYBJC00960].

References

Bachmann, R.; Mizrahi, D.; Atanov, A.; et al. 2022. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision (ECCV)*, 348–367. Springer Nature Switzerland.

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.

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.

Bouchacourt, D.; Tomioka, R.; and Nowozin, S. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Daunhawer, I.; Sutter, T. M.; Chin-Cheong, K.; et al. 2021. On the limitations of multimodal VAEs. ArXiv preprint arXiv:2110.04121.

Daunhawer, I.; Sutter, T. M.; Marcinkevičs, R.; and Vogt, J. E. 2020. Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models. In *DAGM German Conference on Pattern Recognition*, 459–473. Berlin: Springer.

Gonzalez-Garcia, A.; Weijer, J. V. D.; and Bengio, Y. 2018. Image-to-image translation for cross-domain disentanglement. ArXiv preprint arXiv:1805.09730.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

Hinton, G. E. 1999. Products of experts. In *Proceedings of the 9th International Conference on Artificial Neural Networks*, 1–6. Berlin: Springer.

Hirt, M.; Campolo, D.; Leong, V.; et al. 2023. Learning multi-modal generative models with permutation-invariant encoders and tighter variational objectives. ArXiv preprint arXiv:2309.00380.

Hwang, H.; Kim, G.; Hong, S.; et al. 2020. Variational Interaction Information Maximization for Cross-domain Disentanglement. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33. Cambridge, MA: MIT Press.

Jiang, Z.; Zheng, Y.; Tan, H.; et al. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. ArXiv preprint arXiv:1611.05148.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. https://arxiv.org/abs/1312.6114. ArXiv preprint arXiv:1312.6114.

Korthals, T.; Rudolph, D.; Leitner, J.; et al. 2019. Multimodal generative models for learning epistemic active sensing. In *2019 International Conference on Robotics and Automation (ICRA)*, 3319–3325. IEEE.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; et al. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 35–51.

Lee, M.; and Pavlovic, V. 2021. Private-shared disentangled multimodal VAE for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1692–1700.

Liu, H.; Yu, S.; Hou, Q.; et al. 2024. Cross-Modality Image Transformation Using Generative Adversarial Network. In *International Conference on Man-Machine-Environment System Engineering*, 463–468. Springer Nature Singapore.

Mancisidor, R. A.; Kampffmeyer, M.; Aas, K.; et al. 2024. Discriminative multimodal learning via conditional priors in generative models. *Neural Networks*, 169: 417–430.

Meng, D.; Tzelepis, C.; Patras, I.; et al. 2025. MM2Latent: Text-to-facial image generation and editing in GANs with multimodal assistance. In *European Conference on Computer Vision (ECCV)*, 88–106. Springer, Cham.

Mondal, A. K.; Sailopal, A.; Singla, P.; et al. 2023. SSDMM-VAE: variational multi-modal disentangled representation learning. *Applied Intelligence*, 53(7): 8467–8481.

Netzer, Y.; Wang, T.; Coates, A.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2, 4.

Oshima, Y.; Suzuki, M.; Matsuo, Y.; et al. 2024. Enhancing Unimodal Latent Representations in Multimodal VAEs through Iterative Amortized Inference. ArXiv preprint arXiv:2410.11403.

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*. OpenReview.

Palumbo, E.; Laguna, S.; Chopard, D.; et al. 2023. Deep generative clustering with multimodal variational autoencoders.

Palumbo, E.; Manduchi, L.; Laguna, S.; et al. 2024. Deep generative clustering with multimodal diffusion variational autoencoders. The Twelfth International Conference on Learning Representations.

Qiu, P.; Zhu, W.; Kumar, S.; et al. 2025. Multimodal Variational Autoencoder: A Barycentric View. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20060–20068.

Senellart, A.; and Allassonnière, S. 2025. Bridging the inference gap in Multimodal Variational Autoencoders. https://arxiv.org/abs/2502.03952. ArXiv preprint arXiv:2502.03952.

Shi, Y.; Siddharth, N.; Paige, B.; and Kautz, J. 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 15692–15703. Cambridge, MA: MIT Press.

Sutter, T.; Meng, Y.; Agostini, A.; et al. 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. 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.

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

Wang, Q.; Ding, Z.; Tao, Z.; et al. 2021. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wu, M.; and Goodman, N. 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 5580–5590. Cambridge, MA: MIT Press.

Wu, M.; and Goodman, N. 2019. Multimodal generative models for compositional representation learning. ArXiv preprint arXiv:1912.05075.

Xu, G.; Wen, J.; Liu, C.; et al. 2024. Deep variational incomplete multi-view clustering: Exploring shared clustering structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16147–16155.

Yuan, S.; Cui, J.; Li, H.; et al. 2024. Learning Multimodal Latent Generative Models with Energy-Based Prior. In *European Conference on Computer Vision*, 86–100. Springer Nature Switzerland.

Yuan, S.; Lipizzi, C.; and Han, T. 2024. Learning Multimodal Latent Space with EBM Prior and MCMC Inference. ArXiv preprint arXiv:2408.10467.

Zhan, F.; Yu, Y.; Wu, R.; et al. 2023. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4.

Zhang, H.; Koh, J. Y.; Baldridge, J.; et al. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 833–842.

Zhou, D.; Zhang, H.; Ma, J.; et al. 2023. BC-GAN: A Generative Adversarial Network for Synthesizing a Batch of Collocated Clothing. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5): 3245–3259.

Zhu, Z.; Li, Y.; Lyu, W.; et al. 2024. Consistent multimodal generation via a unified GAN framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5048–5057.