Learning Robust Multi-view Representation Using Dual-masked VAEs

Jiedong Wang[1] , **Kai Guo**[1] , **Peng Hu**[1] , **Xi Peng**[1,2] , **Hao Wang**[1*]

[1]College of Computer Science, Sichuan University, China

[2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, China

{wangjd.cs, kaiguo.gm, penghu.ml, pengx.gm, cshaowang}@gmail.com

Abstract

Most existing multi-view representation learning methods assume view-completeness and noise-free data. However, such assumptions are strong in real-world applications. Despite advances in methods tailored to view-missing or noise problems individually, a one-size-fits-all approach that concurrently addresses both remains unavailable. To this end, we propose a holistic method, called Dual-masked Variational Autoencoders (DualVAE), which aims at learning robust multi-view representation. The DualVAE exhibits an innovative amalgamation of dual-masked prediction, mixture-of-experts learning, representation disentangling, and a joint loss function in wrapping up all components. The key novelty lies in the dual-masked (view-mask and patch-mask) mechanism to mimic missing views and noisy data. Extensive experiments on four multi-view datasets show the effectiveness of the proposed method and its superior performance in comparison to baselines. The code is available at https://github.com/XLearning-SCU/2025-IJCAI-DualVAE.

1 Introduction

Many applications face the situation where each data instance in a set $\mathbf{X} = \{x_1, ..., x_n\}$ is sampled from multiple views or even multiple modalities. Here each $x_i|_{i=1}^N$ is denoted by multiple views, e.g., m views $\{x_i^1, ..., x_i^m\}$. Such forms of data are referred to as multi-view data. Multi-view data provide richer and more comprehensive information from raw features within data objectives compared to single-view data [Liang *et al.*, 2024; Yang and Wang, 2018]. As a cutting-edge research topic, multi-view representation learning (MvRL) addresses the motivation of discovering a shared representation from different views with the complex underlying correlation [Zheng *et al.*, 2023; Wang *et al.*, 2015] and later adapting it for downstream tasks, such as human activity recognition [Yadav *et al.*, 2021], 3D reconstruction [Xie *et al.*, 2019] and anomaly detection [Wang *et al.*, 2023].

*Corresponding author (H. Wang).

Over the years, MvRL has showcased promising performance, along with its potential to inspire research in the realm of multi-view or multi-modal AI. MvRL methods primarily focus on learning shared information (i.e., view consistency) among views and distinctive information (i.e., view specificity) within each view [Xu *et al.*, 2021; Ke *et al.*, 2024]. Most of them assume that all views are complete and data are noise-free, involving a conditional scenario that we call *closed multi-view setting*. However, the closed multi-view setting is too limited for real-world applications as it frequently happens that some data instances are not sampled in certain views and data features are corrupted by noise, which refers to a more realistic scenario that we call *generic multi-view setting*. This paper is concerned with view-missing and sample-noise problems in such a generic setting.

View-missing and sample-noise would disturb multi-view data, and sequentially impact the data's benefit for downstream tasks. To date, some efforts have been devoted to solving view-missing issues [Zhang *et al.*, 2018; Wen *et al.*, 2019; Zhang *et al.*, 2020] and devising noise-tolerance multi-view models [Yue *et al.*, 2019]. It is worth noting that the existing methods can only tackle view missingness or noise robustness, rather than addressing both of them within a unified framework. That is, the robustness of a unified model for both cases is still challenging and non-trivial in practice. Specifically, there are three major challenges: 1) *view representation*, 2) *view fusion*, and 3) *view disentangle*, due to the problems of view-missing and noisy data.

As expatiated above, there is no existing work that can address all issues holistically using a single model in the generic multi-view setting. In this work, we aim to deal with view-missing and sample-noise problems using a holistic model. To this end, we propose Dual-masked Variational Autoencoders (DualVAE), which jointly aim at learning robust view-consistent and view-specific representations. Our main ideas here are three-fold: *dual-masked prediction*, *mixture-of-experts learning*, and *representation disentangling*. We will expatiate on each of them shortly. In a nutshell, the pipeline of the proposed DualVAE is as follows: i) view-specific encoders take the input of each view and learn view-specific representation against view-specific noise within each view, ii) a consistent encoder with dual-masked prediction and mixture-of-experts learning for processing all views, yielding robust view-consistent representations, and iii) a disentangled mod-

ule that minimizes the upper-bound of mutual information to disentangle view-consistent and view-specific representations such that the distribution of view-consistency can be maximized to differ from the view-specific information of each view, thereby extracting robust shared information.

In summary, we make the following contributions:

- We delve into multi-view representation learning in a generic multi-view setting, and propose a robust multi-view representation learning method called DualVAE to solve both view-missing and sample-noise problems.
- We devise a novel dual-masked mechanism, together with a Mixture-of-Experts layer and a disentangle learning module in addressing the distinct challenges posed by view-missing and sample-noise in multi-view data.
- Experimental results using four real-world datasets demonstrate the proposed DualVAE outperforms baseline methods dramatically.

2 Related Work

Multi-view Representation Learning. The objective of multi-view representation learning (MvRL) is to extract high-quality feature representations from diverse sources of data for downstream tasks [Wang *et al.*, 2015; Chen *et al.*, 2022; Yacobi *et al.*, 2024]. In recent years, exciting progress has been achieved in MvRL by assuming view-completeness and noise-free multi-view data [Hwang *et al.*, 2021; Xu *et al.*, 2022; Ke *et al.*, 2024]. However, MvRL encounters the challenge of poor quality on input data such as view-missing and sample-noise [Li *et al.*, 2018]. As for their specific objectives, the methods dedicated to improving model robustness can be generally categorized into the following two types: 1) incomplete MvRL methods that aim at learning view-common information across observable views to handle incomplete information such as partially aligned or partially missingness in multi-view data [Zhang *et al.*, 2018; Wen *et al.*, 2019; Zhang *et al.*, 2020], and 2) noise-robust MvRL methods that focus on employing specific methods like filtering noises and disentangling noises to address data noises in different views [Yue *et al.*, 2019; Fan *et al.*, 2023; Wang *et al.*, 2025].

It is worth noting that existing MvRL methods rarely consider a generic multi-view setting where the data encounter both view-missingness and sample-noise problems. Our approach is a holistic method that can handle both of them.

Multi-view VAE. With the advancement of generative models, variational autoencoders (VAE) technologies have been introduced into the realm of multi-view research, called multi-view VAE. Multi-view VAE typically extracts generalizable representations by learning a joint distribution of multi-view data [Daunhawer *et al.*, 2021; Sutter *et al.*, 2021]. Following Wu and Goodman [2018], existing multi-view VAE methods show promising results for learning high-quality representations across multiple views, which is beneficial for downstream tasks such as clustering and classification [Shi *et al.*, 2019; Sutter *et al.*, 2020; Sutter *et al.*, 2021].

Some efforts have also been devoted to learning latent representations from the subsets of observable views to address missing views [Wu and Goodman, 2018; Vedantam *et al.*, 2017; Shi *et al.*, 2019; Sutter *et al.*, 2021]. However, their computational cost increases significantly as the combinatorial number of views increases, which leads to a decrease in efficiency. Our approach adopts a dual-masked mechanism, which deals with both view-missing and sample-noise problems without increasing such combinatorial computing cost.

Masked VAE. Masking technique is a self-supervised learning (SSL) method that randomly masks some partial information (e.g., words in a sentence or patches in an image) and then prompts the model to predict invisible blocks, which receives a lot of attention in NLP and CV community [Vaswani, 2017; Devlin, 2018; Bao *et al.*, 2021; He *et al.*, 2022]. The masking mechanism has also been incorporated into VAE, resulting in the development of masked VAE [Xia *et al.*, 2023; Li *et al.*, 2023; Xu *et al.*, 2023] and multi-view masked VAE [Ke *et al.*, 2024].

Multi-view masked VAE [Ke *et al.*, 2024] uses pixel-masking for input images and does not consider view-missing and sample-noise problems in the generic multi-view setting. Our approach is a dual-masked VAE, which consists of view-mask and patch-mask, along with mixture-of-experts learning and representation disentangling, an innovative amalgamation to enhance model robustness against both view-missing and sample-noise problems.

3 Methodology

As aforementioned, we propose a DualVAE for multi-view representation learning. Prior to introducing the details of the model, we first clarify the problem studied in this paper.

Definition 1 (Problem Statement)**.** *Given a multi-view dataset with m views and n samples* $\mathbf{X} = \{x_i | x_i^1, ..., x_i^m\}_{i=1}^n$ *in a generic setting, where some data might have unavailable views (e.g., j-view) and corrupted parts (e.g.,* $x_{i,k}^j$ *where k denotes dimension), referred to as missing views and noisy samples respectively, the dataset is used to train a model. The trained model is then employed to drive multi-view representations for downstream tasks after deployment, where the test may also contain view-missing and sample-noisy data.*

3.1 Overall Architecture

Figure 1 illustrates the architecture of our DaulVAE, comprising three novel components: i) *Dual-masked Prediction* (DMP), which processes multi-view data using view-mask and patch-mask, ii) *Mixture-of-Experts* (MoE), which models view-consistent representation, and iii) *Representation Disentangling*, which disentangles view-consistent and view-specific representations. Specifically, we first employ a set of view-specific encoders $\{E_s^i\}_{i=1}^m$ to extract view-specific representation $\{\mathbf{s}^i\}_{i=1}^m$ which contains noises from corresponding views. We also perform data augmentation on the original multi-view data by cropping images into blocks. Then we use the DMP to mask views and feature patches. Next, those visible blocks and views are fed into a view-shared encoder E_c to learn a posterior distribution of the latent space $\{q(\mathbf{c}|x^i)\}$ from different views. To mitigate the impact of missing views, we aggregate these distributions by using a

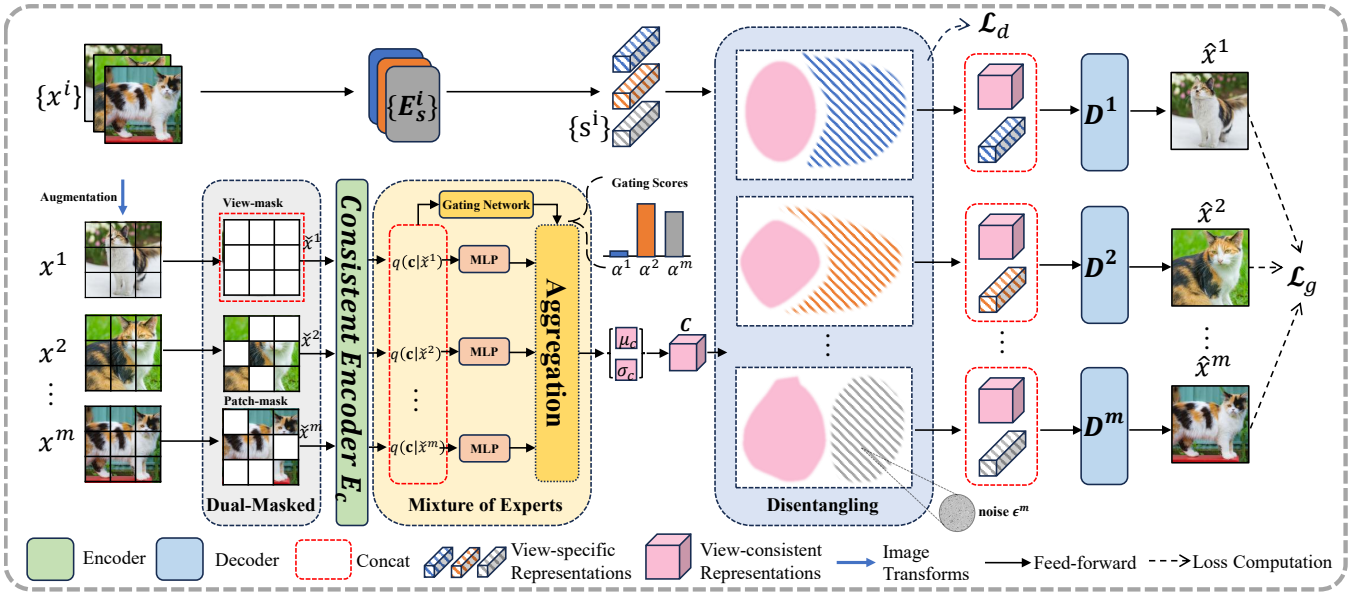


Figure 1: An overview of the proposed DualVAE. To clarify, we use a set of multi-view data with three views for illustration. The goal of this model is to extract view-consistent representation $\mathbf{c}$ in the generic multi-view setting for downstream tasks. DualVAE mainly consists of three modules: Dual-masked Prediction (*gray* block), Mixture-of-Experts (*yellow* block), and Disentangling Learning (*blue* block). The dual-masked prediction module processes multi-view data by performing view-mask and patch-mask. Mixture-of-Experts aggregates posteriors learned by the view-shared encoder. Gating scores are generated by the concatenation of posteriors to mitigate the impact of missing views. Representation disentangling module separates view-consistency and view-specificity which contain sample-noise. Finally, DualVAE utilizes view-consistency and view-specificity to generate data. [Best viewed in color]

MoE layer. In practice, we reparameterize from the aggregated distribution to obtain view-consistent representation $\mathbf{c}$. Subsequently, we use a disentangling module by minimizing the mutual information (MI) between $\mathbf{c}$ and $\{\mathbf{s}^i\}_{i=1}^m$ to separate representations from noise and enhance the quality of $\mathbf{c}$. Finally, we concatenate $\mathbf{c}$ and $\{\mathbf{s}^i\}_{i=1}^m$ and feed the results into decoders to reconstruct multi-view data.

Roadmap to Our Model. Given the problem and architecture, our primary objectives for designing DualVAE are (1) *processing multi-view data* using DMP (**Section 3.2**), (2) *modeling view-consistent representation* with MoE (**Section 3.3**), and (3) *learning disentangled representations* using upper-bound MI (**Section 3.4**). We now elaborate on our solutions to each component.

3.2 Dual-masked Prediction

In this work, we aim to extract robust multi-view representations in the generic settings. We denote that the denoising autoencoders in the domain of computer vision apply different levels of distortion to the input images and then prompt the model to recover the degraded areas of the images [Vincent *et al.*, 2008]. Following [Vincent *et al.*, 2008], we devise a view-mask that corrupts multi-view data by discarding one or more views of the data to enhance robustness against view-missing. Moreover, noisy samples in the data would degrade a model in extracting shared information among multi-view data. To address this challenge, we use a patch-mask, which has demonstrated its robustness against sample-noise in [He *et al.*, 2022; Xia *et al.*, 2023] by randomly masking tokens and reconstructing them later. That is, our masking approach is a dual-masked model of *view-mask* and *patch-mask*.

Specifically, view-mask focuses on the adaptability to view-missing data. In practice, we randomly select samples $\{\mathbf{X}_{k_1}, \mathbf{X}_{k_2}, \ldots, \mathbf{X}_{k_n}\} \subseteq \{\mathbf{X}_i\}_{i=1}^n$. For each sample, we next randomly select one view (e.g., the v-th view) and then mask its data $x_{k_i}^v$. An intuition of our view-mask is that it simulates view-missing scenarios and prompts the model to predict invisible views so as to enhance robustness against view-missing. Patch-mask focuses on the noise tolerance of the model. Since noise-corrupted views share a common latent space, our motivation here is that the patch-mask has capacity to shield most noise and force the model to extract more information gains from data. By leveraging these visible blocks as inputs to the view-shared encoder E_c, the model then predicts the original information of incomplete data. We call such a method *Dual-masked Prediction* (DMP), formulated as $\check{\mathbf{X}} = DMP(\mathbf{X})$. By training the view-shared encoder, the data of each view are mapped to the shared latent space, denoted as posterior distributions $\{q_\theta(\mathbf{c}|\check{x}^i)\}$, where $\check{x}^i$ is the output of dual-masked module on x^i and θ is trainable parameters of the view-shared encoder. Moreover, our dual-masked model processes multi-view data without incurring additional computational overhead.

3.3 Mixture-of-Experts Learning

To extract view-consistent representation, a common solution is to aggregate information from multiple views, referring to *view-fusion*. Since the existence of missing views and noisy samples in multi-view data, view fusion is challenging in

generic settings. Aiming to extract coherent representations in terms of the shared latent spaces, we propose to mitigate the interference caused by missing views. Our main idea is to reduce the weight coefficient of the distribution from missing views. To this end, we saddle the model with a Mixture-of-Experts (MoE) layer.

Specifically, as the missingness of samples and nonidentically distributed noises from multiple views, the posterior $q(\mathbf{c}|\check{x}^i)$ generated by the view-shared encoder might not be optimal. To address this challenge, we assign each view an expert (as shown in Figure 1), to process the posterior from each view, namely, transforming the posterior and further learning on the transition space to the shared latent space. In our framework, each expert is implemented by a Multilayer Perceptron (MLP) parameterized by ϕ^i. The transformed posteriors are expressed as $q_{\phi^i}(\mathbf{c}|\check{x}^i) = MLP_{\phi^i}(q_\theta(\mathbf{c}|\check{x}^i))$. Then we have

$$\begin{aligned} q_\phi(\mathbf{c}|\check{\mathbf{X}}) &= MoE(\{q_\theta(\mathbf{c}|\check{x}^i)\}) \\ &\propto \sum \alpha^i \cdot q_{\phi^i}(\mathbf{c}|\check{x}^i) \end{aligned} \tag{1}$$

where $\check{\mathbf{X}} = DMP(\mathbf{X})$. $\boldsymbol{\alpha} = \{\alpha^i\}_{i=1}^m$ are gating scores generated by a gating network as below

$$\boldsymbol{\alpha} = Softmax(Gate_\gamma(\{q_\theta(\mathbf{c}|\check{x}^i)\}_{i=1}^m)) \tag{2}$$

where γ denotes the parameters of the gating network. The intuition here is that by dynamically reducing gating scores α_i of the missing views, the model can mitigate their impact to improve the quality of view-consistent representations for view-missing and sample-noise data. In addition, to further enhance the model's adaptability, we use a normal Gaussian distribution to model the prior of view-consistent representation, formulated as $p(\mathbf{c}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Moreover, as the random sampling from posterior $q_\phi(\mathbf{c}|\check{\mathbf{X}})$ cannot directly propagate gradients, we reparameterize this distribution to obtain view-consistent representation $\mathbf{c}$, formulated as follows

$$q_\phi(\mathbf{c}|\check{\mathbf{X}}) = \mathcal{N}(\mu_\mathbf{c}, {\sigma_\mathbf{c}}^2) = \mu_c + \epsilon_c \sigma_c \tag{3}$$

where μ_c and σ_c are trainable parameters and $\epsilon_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.4 Representation Disentangling

Due to the diversity of multi-view data, the noise distributions in different views are also distinct. As the noises are often irregular and random, enhancing the noise-tolerance of MvRL is challenging and non-trivial. To address this challenge, our insight is that view-specific representation might contain both view-specific features and sample noise, as shown below

$$p(\mathbf{s}^i) = \xi(s^i, \epsilon^i) \tag{4}$$

where s^i denotes view-specific information, ϵ^i denotes sample-noise, and ξ denotes a joint distribution of them. That is, multi-view data contains three types of information: view-consistent information, view-specific information, and distinct noise. As shown in Section 3.2, we proposed a novel dual-masked prediction to improve the quality of view-consist representation, where the masked content might be view-specific information or noisy data. We now introduce how to disentangle them.

As mentioned above, suppose we have extracted a view-consistent representation, we then propose to explore the remaining information across views. Our motivation is that view-consistency and view-specificity should be independent. We design a disentangle module by minimizing the upper-bound of mutual information between them. Specifically, as illustrated in Figure 1, we deploy a set of view-specific encoders $\{E_s^i\}_{i=1}^m$ to tackle each views, which extract view-specific representation $\{\mathbf{s}^i\}_{i=1}^m$. We formulate the posterior distributions of $\mathbf{s}^i$ as $q_{\varphi^i}(\mathbf{s}^i|x^i)$, where φ^i denotes trainable parameters of E_s^i. Similarly, we employ a Gaussian distribution to model view-specific representation, formulated as $p(\mathbf{s}^i) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Then, the view-consistent representation $\mathbf{c}$ and view-specific representation $\{\mathbf{s}^i\}_{i=1}^m$ are the inputs of disentangling module. After that, we concatenate view-consistent representation and view-specific representation, denoted as $\mathbf{z}^i = [\mathbf{c}, \mathbf{s}^i]$. Finally, we deploy a set of decoders $\{D^i\}_{i=1}^m$ to generate data, where $\mathbf{z}^i$ is the input of the decoder D^i.

Based on Eq. (4) and the chain rule for mutual information, we have

$$\begin{aligned} I(\mathbf{s}^i; \mathbf{c}) &= I((\epsilon^i, s^i); \mathbf{c}) \\ &= I(\epsilon^i; \mathbf{c}) + I(s^i; \mathbf{c}|\epsilon^i) \\ &\geq I(\epsilon^i; \mathbf{c}) \end{aligned} \tag{5}$$

which shows that we can disentangle $\mathbf{c}$ and ϵ^i by minimizing the upper bound of mutual information between $\mathbf{c}$ and $\mathbf{s}^i$ to address noise data. The mutual information between $\mathbf{s}^i$ and $\mathbf{c}$ is formulated as follows

$$\begin{aligned} I(\mathbf{s}^i, \mathbf{c}) &= \mathbf{E}_{p(\mathbf{s}^i, \mathbf{c})}[\log \frac{p(\mathbf{s}^i|\mathbf{c})}{p(\mathbf{s}^i)}] \\ &\leq \mathbf{E}_{p(\mathbf{s}^i, \mathbf{c})}[\log \frac{p(\mathbf{s}^i|\mathbf{c})}{h(\mathbf{s}^i)}] \\ &= \mathbf{KL}(p(\mathbf{s}^i|\mathbf{c})||h(\mathbf{s}^\mathbf{i})) \end{aligned} \tag{6}$$

where $h(\mathbf{s}^i)$ denotes the variational marginal approximation of $\mathbf{s}^i$. However, the upper bound of Eq. (6) is difficult to estimate. Following CLUB [Cheng *et al.*, 2020], we use an approximation without assuming a prior. CLUB approximates $p(\mathbf{s}^i|\mathbf{c})$ using a variational distribution $q_{\psi^i}(\mathbf{s}^i|\mathbf{c})$, where ψ^i denotes the parameters of the i-th estimator. Then, we define our disentangling loss function as

$$\mathcal{L}_d = \sum_{i=1}^m \mathcal{L}_d^i \tag{7}$$

where $\mathcal{L}_d^i$ is the loss of the i-th estimator, formulated as

$$\begin{aligned} \mathcal{L}_d^i =& I_{CLUB}(\mathbf{s}^i; c) \\ =& \mathbf{E}_{p(\mathbf{s}^i, \mathbf{c})}[\log q_{\psi^i}(\mathbf{s}^i|\mathbf{c})] - \\ & \mathbf{E}_{p(\mathbf{c})}\mathbf{E}_{p(\mathbf{s}^i)}[\log q_{\psi^i}(\mathbf{s}^i|\mathbf{c})] \end{aligned} \tag{8}$$

where $I_{CLUB}(\mathbf{s}^i; \mathbf{c}) \geq I(\mathbf{s}^i, \mathbf{c})$ as proved in [Cheng *et al.*, 2020].

We then concatenate view-consistent representation and view-specific representations to generate $\mathbf{z}^i$, i.e., $\mathbf{z}^i = [\mathbf{c}, \mathbf{s}^i]$,

whose posterior is defined as $q_{\phi,\varphi^i}(\mathbf{z}^{\mathbf{i}}|\{x^i\})$. Similar to view-consistent representation, we yet reparameterize the posteriors of view-specific representation as follows

$$q_{\varphi^i}(\mathbf{s}^i|x^i) = \mathcal{N}(\mu_s^i, (\sigma_s^i)^2) = \mu_s^i + \epsilon_s^i \sigma_s^i \tag{9}$$

where $\{\mu_{s^i}\}$, and $\{\sigma_{s^i}\}$ are trainable parameters in neutral networks and $\epsilon_s^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The sampling results from the latent distribution are then fed into decoders to reconstruct samples denoted as $\{\hat{x}^i\}_{i=1}^m$. The likelihood of the reconstructed samples of i-th view is shown as follows

$$\hat{x}^i = p(x^i|\mathbf{z}^i). \tag{10}$$

In practice, we adopt vanilla VAE to build a base model [Kingma, 2013]. The reconstruction loss $\mathcal{L}_g$ with the evidence lower bound (ELBO) can be expressed as:

$$\begin{aligned}\mathcal{L}_{ELBO}(x^i) =& \mathbf{E}_{q_{\phi,\varphi^i}(\mathbf{z}^i|\{x^i\})}[\log p(x^i|\mathbf{z}^i)] - \\ & \mathbf{KL}[q_{\phi,\varphi^i}(\mathbf{z}^i|\{x^i\})\|p(\mathbf{z}^i)].\end{aligned} \tag{11}$$

Suppose $\mathbf{c}$ and $\mathbf{s}^i$ are conditionally independent, then we have $q(\mathbf{z}^{\mathbf{i}}|\{x^i\}) = q(\mathbf{c}, \mathbf{s}^i|\{x^i\}) = q(\mathbf{c}|\{x^i\})q(\mathbf{s}^i|x^i)$. The KL divergence in Eq. (11) is then formulated below

$$\begin{aligned}&\mathbf{KL}\left[q\left(\mathbf{z}^i|\{x^i\}\right)\middle\|p\left(\mathbf{z}^i\right)\right] = \mathbf{KL}\left[q\left(\mathbf{c}, \mathbf{s}^i|\{x^i\}\right)\middle\|p\left(\mathbf{c}, \mathbf{s}^i\right)\right] \\ &= \mathbf{E}_{q(\mathbf{c},\mathbf{s}^i|\{x^i\})}\left[\log \frac{q\left(\mathbf{c}, \mathbf{s}^i|\{x^i\}\right)}{p\left(\mathbf{c}, \mathbf{s}^i\right)}\right] \\ &= \mathbf{E}_{q(\mathbf{c}|\{x^i\})}\mathbf{E}_{q(\mathbf{s}^i|x^i)}\left[\log \frac{q\left(\mathbf{c}|\{x^i\}\right)q\left(\mathbf{s}^i|x^i\right)}{p\left(\mathbf{c}\right)p\left(\mathbf{s}^i\right)}\right] \\ &= \mathbf{E}_{q(\mathbf{c}|\{x^i\})}\mathbf{E}_{q(\mathbf{s}^i|x^i)}\left[\log \frac{q\left(\mathbf{c}|\{x^i\}\right)}{p\left(\mathbf{c}\right)} + \log \frac{q\left(\mathbf{s}^i|x^i\right)}{p\left(\mathbf{s}^i\right)}\right] \\ &\approx \mathbf{KL}\left[q\left(\mathbf{c}|\{\check{x}^i\}\right)\middle\|p\left(\mathbf{c}\right)\right] + \mathbf{KL}\left[q\left(\mathbf{s}^i|x^i\right)\middle\|p\left(\mathbf{s}^i\right)\right].\end{aligned} \tag{12}$$

However, $\mathbf{KL}[q(\mathbf{c}|\{\check{x}^i\})||p(\mathbf{c})]$ cannot be computed directly. By Eq. (1) and dropping ϕ^i for simplicity, we have

$$\begin{aligned}\mathbf{KL}(q(\mathbf{c}|\check{\mathbf{X}})\|p(\mathbf{c})) &= \mathbf{E}_{q(\mathbf{c}|\check{\mathbf{X}})}\left[\log \frac{p(\mathbf{c}|\check{\mathbf{X}})}{p(\mathbf{c})}\right] \\ &\propto \mathbf{E}_{q(\mathbf{c}|\check{\mathbf{X}})}\left[\log \frac{\sum \alpha^i \cdot q(\mathbf{c}|\check{x}^i)}{p(\mathbf{c})}\right].\end{aligned} \tag{13}$$

By Jensen's inequality, we can derive

$$\begin{aligned}\mathbf{E}_{q(\mathbf{c}|\check{\mathbf{X}})}\left[\log \frac{\sum \alpha^i \cdot q(\mathbf{c}|\check{x}^i)}{p(\mathbf{c})}\right] &\geq \sum_{i=1}^m \alpha^i \cdot \mathbf{E}_{q(\mathbf{c}|\check{x}^i)}\left[\log \frac{q(\mathbf{c}|\check{x}^i)}{p(\mathbf{c})}\right] \\ &= \sum_{i=1}^m \alpha^i \cdot \mathrm{KL}(q(\mathbf{c}|\check{x}^i)\|p(\mathbf{c})).\end{aligned} \tag{14}$$

Then, the ELBO of the model can be reformulated as

$$\begin{aligned}\mathcal{L}_{ELBO} &= \sum \mathcal{L}_{ELBO}(x^i) \\ &\propto \sum_{i=1}^m \mathbf{E}_{q_{(\mathbf{z}^i|\{x^i\})}}[\log p(x^i|\mathbf{z}^i)] - \\ &\quad \sum_{i=1}^m \mathbf{KL}\left[q\left(\mathbf{s}^i|x^i\right)\|p\left(\mathbf{s}^i\right)\right] - \\ &\quad \mathbf{KL}(q(\mathbf{c}|\{\check{x}^i\})\|p(\mathbf{c}))\end{aligned} \tag{15}$$

Objection Function. Formally, we have the total loss function of our DualVAE:

$$\mathcal{L}_{loss} = \mathcal{L}_{ELBO} - \mathcal{L}_d. \tag{16}$$

4 Experiments

4.1 Experimental Setup

Datasets. To evaluate the proposed model, we conduct experiments on four publicly available multi-view datasets. The statistics of each dataset are shown in Table 1.

Dataset	Samples	Views	Classes
COIL-20 [Nene *et al.*, 1996b]	1,440	3	20
COIL-100 [Nene *et al.*, 1996a]	7,200	3	100
E-MNIST [Liu and Tuzel, 2016]	70,000	2	10
PolyMNIST [Palumbo *et al.*, 2023]	70,000	5	10

Table 1: Data statistics of the benchmark datasets.

Baselines. We compare DualVAE against the following three types of baseline methods:

- Single-view VAE: β-VAE [Higgins *et al.*, 2017], and Joint-VAE [Dupont, 2018].
- Multi-view non-VAE: SCM [Luo *et al.*, 2024], and MFLVC [Xu *et al.*, 2022].
- Multi-view VAE: MVAE [Wu and Goodman, 2018], Multi-VAE [Xu *et al.*, 2021], MIB [Federici *et al.*, 2020], and MRDD [Ke *et al.*, 2024].

For those sing-view VAE methods, we choose the best results among all views. For the multi-view VAE methods that are tailored for two views, we select the first two views as inputs. For other baseline methods, we ran their original systems with the settings as they suggested.

Implementation Details. We deployed the proposed model and baselines using Pytorch 2.3.1 and ran experiments on NVIDIA TITAN GPUs with 24GB of memory in Ubuntu 20.04.1 LTS. We utilize Adam optimizer with weight decay and set training epochs as 200. Moreover, we set the initial learning rate to 1×10^{-4} and adopt cosine annealing during the training. Following [He *et al.*, 2022; Ke *et al.*, 2024], the default patch-mask ratio is 0.7. We then initially set view-mask ratio as 0.3. Both the dimensions of view-consistent representation and view-specific representation are 10. We ran 10 times on each evaluation and recorded the average value and standard deviation.

4.2 Experimental Results

Task Settings. We evaluate each model with clustering and classification tasks. For clustering, we implement K-means algorithm on the learnt representations. For classification, we employ support vector classification (SVC). To have a comprehensive evaluation, we evaluate each model under different settings of *View-Missing Ratio* (VMR) and *Sample-Noise Ratio* (SNR). The VMR is defined as $VMR = k/N$, where N is the size of all samples and k denotes the number of samples randomly selected to remove one of the views. For SNR, we

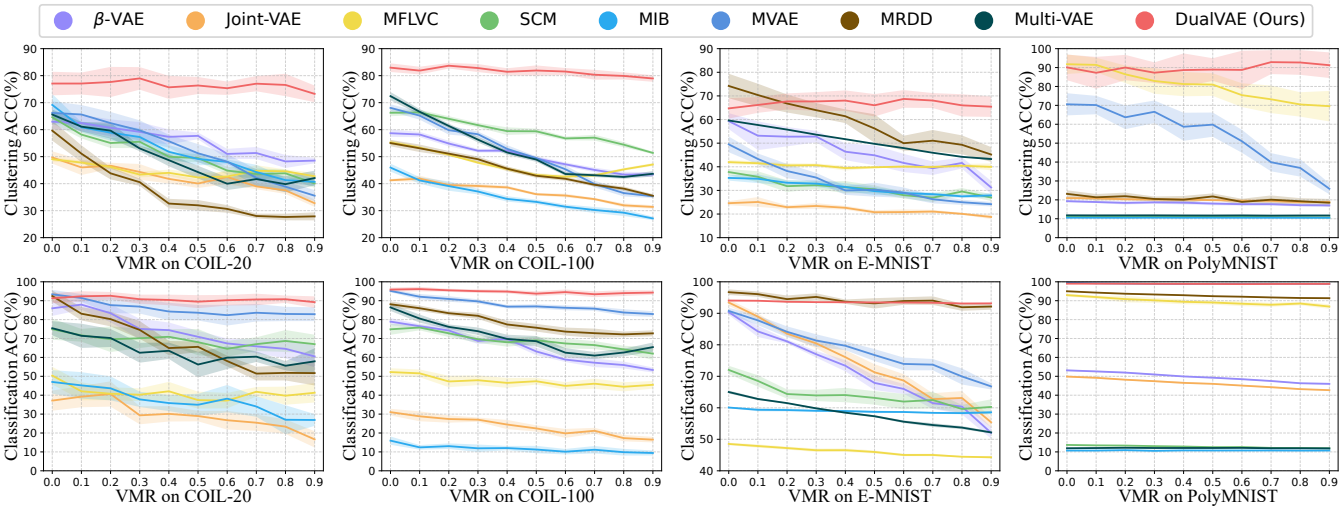


Figure 2: Clustering & Classification performance comparison with different View-Missing Ratio (VMR) settings on four datasets.

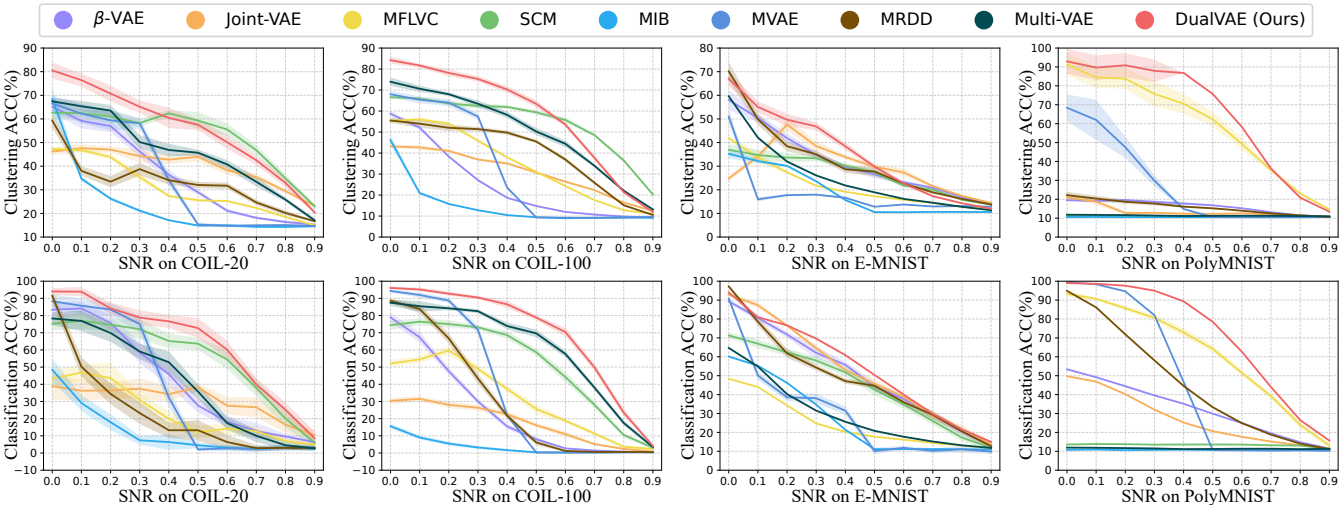


Figure 3: Clustering & Classification performance comparison with different Sample-Noise Ratio (SNR) settings on four datasets.

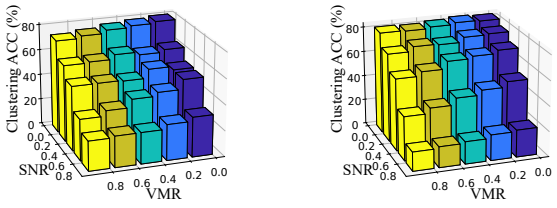
added salt-and-pepper noise into four datasets with varying degrees of intensity, denoting the ratio of pixels that are corrupted by noises. we vary the values of VMR and SNR from 0% to 90%, with an interval of 10%.

We use Accuracy (ACC) as the metric to evaluate the performance of each model for both clustering and classification tasks on four datasets.

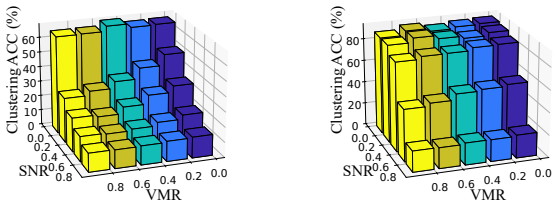
Overall Evaluation. The evaluation results in terms of ACC on each dataset with different VMR and SNR settings are shown in Figure 2 and Figure 3. From the results, we have the following observations:

- The proposed DualVAE is superior to baseline methods and achieves promising performance even on the heavy missing settings. DualVAE performs well with the increasing of VMR. The results demonstrate that our DualVAE has capability to handle view missingness.
- All the baseline models suffer performance degradation as the SNR increases. In contrast, DualVAE exhibits significant robustness and achieves superior performance.
- The results in the above two experiments clearly show that the proposed DualVAE model is effective against view-missing and sample-noise problems.

In addition, to further demonstrate the performance of the proposed DualVAE in a generic setting, we examine the model under a mixed setting of VMR and SNR, where both variables range from 0% to 80% with an interval of 20%. The results are shown in Figure 4. From the figure, we can see that for each fixed SNR, the model has small variation as VMR increases, which demonstrates its robustness against view-missing problems. The model also exhibits robust performance against noise. The results show that our DualVAE is robust in generic multi-view settings.



(a) COIL-20 (b) COIL-100



(c) E-MNIST (d) PolyMNIST

Figure 4: Clustering performance with different View-Missing Ratio (VMR) and Sample-Noise Ratio (SNR) settings on four datasets.

Ablation Study. To explore the framework of DualVAE, we conduct an ablation study by analyzing the effectiveness of the devised DMP (Section 3.2), MoE (Section 3.3), and disentanglement (Section 3.4). We remove each of them from the framework and conduct clustering tasks on four datasets with the setting of *VMR*=0.5 and *SNR*=0.1. Particularly, we exclude MoE by replacing linear layers with it. Experimental results of the ablation study are shown in Table 2. From the results, we see that the removal of any component leads to a degradation in performance. This indicates that all components in our framework are essential. As for the three components, DMP contributes more than the other two.

Methods	COIL-20	COIL-100	E-MNIST	PolyMNIST
w/o DMP	41.09	46.42	28.92	86.68
w/o MoE	65.57	52.96	40.17	81.35
w/o $\mathcal{L}_d$	63.52	55.48	38.01	64.96
DualVAE	**72.13**	**76.10**	**41.62**	**92.98**

Table 2: **Ablation study on the components of DualVAE**. ACC in clustering with *VMR*=0.5 and *SNR*=0.1. Best scores are in **bold**.

Parameter Study. Since the proposed dual-masked model contains two hyperparameters, i.e., *view-mask ratio* and *patch-mask ratio*, we now explore how they influence the performance of the model. We conduct grid search experiments in the setting of *VMR*=0.5 and *SNR*=0.1, where view-mask ratio and patch-mask ratio vary from 0% to 90% with a 10% interval. First, we conduct classification on four datasets with different view-mask ratios by fixing the patch-mask ratio as 70%. The results are shown in Figure 5. It is worth noting that a proper view-mask ratio can significantly enhance the performance of the model, such as 30%. Then we fix the view-mask ratio as 0.3, and vary the patch-mask ratio for further exploration. The results are shown in Figure 6. From this figure, we can see that a higher patch-mask ratio may

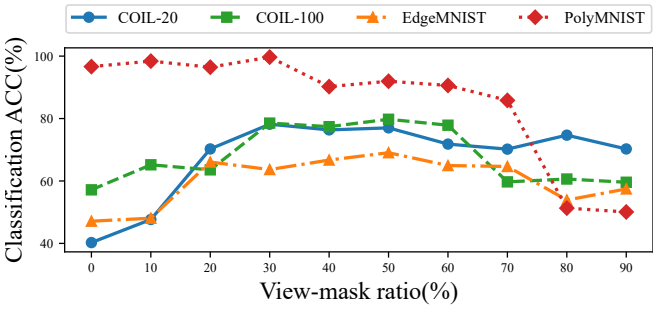


Figure 5: **Parameter study on view-mask ratio.** ACC in classification with the setting of *VMR*=0.5 and *SNR*=0.1.

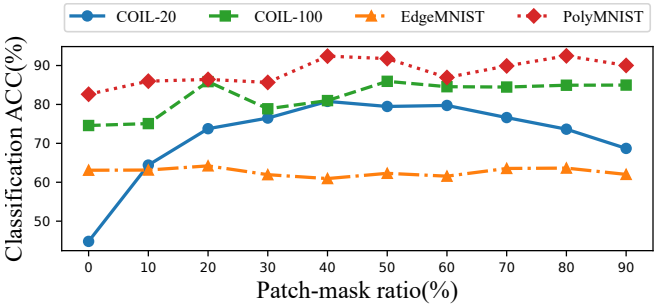


Figure 6: **Parameter study on patch-mask ratio.** ACC in classification with the setting of *VMR*=0.5 and *SNR*=0.1.

achieve better performance. In addition, for small datasets such as E-MNIST and COIL-20, a high-intensity patch-mask would lead to the loss of representative information, resulting in performance degradation. That is, both view-mask and patch-mask contribute to the robustness of multi-view representations learned by the proposed model.

5 Conclusion

In this paper, we are concerned with the robustness of multi-view representation learning in a generic multi-view setting, particularly for view-missing and sample-noise problems. To this end, we proposed a novel Dual-masked Variational AutoEncoders (denoted as DualVAE), which is a unified framework to extract robust representations for multi-view data. The DualVAE exhibits an innovative amalgamation of dual-masked prediction, mixture-of-experts learning, representation disentangling, and a joint loss function in wrapping up all components. Extensive experimental results demonstrate the superior robustness of the proposed method compared to baseline methods in the settings of different view-missing ratios and sample-noise ratios.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (grant no. 2024YFB4710604), NSFC (nos. 62406209, 62472295, U24B20174, and U21B2040), Sichuan Science and Technology Program (nos. 2024NSFTD0130, 2024NSFTD0038, and 2025ZNSFSC1486), and the Fundamental Research Funds for the Central Universities (nos. CJ202303, CJ202403, and YT202421).

References

[Bao *et al.*, 2021] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[Chen *et al.*, 2022] Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241, 2022.

[Cheng *et al.*, 2020] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pages 1779–1788. PMLR, 2020.

[Daunhawer *et al.*, 2021] Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021.

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

[Dupont, 2018] Emilien Dupont. Learning disentangled joint continuous and discrete representations. *Advances in Neural Information Processing Systems*, 31, 2018.

[Fan *et al.*, 2023] Zizhu Fan, Yijing Huang, Chao Xi, and Qiang Liu. Multi-view adaptive k-nearest neighbor classification. *IEEE Transactions on Artificial Intelligence*, 2023.

[Federici *et al.*, 2020] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[Hwang *et al.*, 2021] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34:12194–12207, 2021.

[Ke *et al.*, 2024] Guanzhou Ke, Bo Wang, Xiaoli Wang, and Shengfeng He. Rethinking multi-view representation learning via distilled disentangling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26774–26783, 2024.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Li *et al.*, 2018] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

[Li *et al.*, 2023] Jingxiong Li, Sunyi Zheng, Zhongyi Shui, Shichuan Zhang, Linyi Yang, Yuxuan Sun, Yunlong Zhang, Honglin Li, Yuanxin Ye, Peter MA Van Ooijen, et al. Masked conditional variational autoencoders for chromosome straightening. *IEEE Transactions on Medical Imaging*, 2023.

[Liang *et al.*, 2024] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[Liu and Tuzel, 2016] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[Luo *et al.*, 2024] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4697–4705, 2024.

[Nene *et al.*, 1996a] Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia Object Image Library (COIL-100). In *Technical Report, Department of Computer Science, Columbia University CUCS-006-96*, 1996.

[Nene *et al.*, 1996b] Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia Object Image Library (COIL-20). In *Technical Report, Department of Computer Science, Columbia University CUCS-005-96*, 1996.

[Palumbo *et al.*, 2023] Emanuele Palumbo, Imant Daunhawer, and Julia E Vogt. Mmvae+: Enhancing the generative quality of multimodal vaes without compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview, 2023.

[Shi *et al.*, 2019] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multimodal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[Sutter *et al.*, 2020] Thomas Sutter, Imant Daunhawer, and Julia Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in Neural Information Processing Systems*, 33:6100–6110, 2020.

[Sutter *et al.*, 2021] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*, 2021.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Vedantam *et al.*, 2017] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extract-

ing and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092. PMLR, 2015.

[Wang *et al.*, 2023] Hao Wang, Zhi-Qi Cheng, Jingdong Sun, Xin Yang, Xiao Wu, Hongyang Chen, and Yan Yang. Debunking free fusion myth: Online multi-view anomaly detection with disentangled product-of-experts modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3277–3286, 2023.

[Wang *et al.*, 2025] Xuzheng Wang, Shiyang Lan, Zhihao Wu, Wenzhong Guo, and Shiping Wang. Multi-view representation learning with decoupled private and shared propagation. *Knowledge-Based Systems*, page 112956, 2025.

[Wen *et al.*, 2019] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Hong Liu. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5393–5400, 2019.

[Wu and Goodman, 2018] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[Xia *et al.*, 2023] Yinghao Xia, Changfang Chen, Minglei Shu, and Ruixia Liu. A denoising method of ecg signal based on variational autoencoder and masked convolution. *Journal of Electrocardiology*, 80:81–90, 2023.

[Xie *et al.*, 2019] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2690–2698, 2019.

[Xu *et al.*, 2021] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9234–9243, 2021.

[Xu *et al.*, 2022] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[Xu *et al.*, 2023] Rui Xu, Le Hui, Yuehui Han, Jianjun Qian, and Jin Xie. Scene graph masked variational autoencoders for 3d scene generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5725–5733, 2023.

[Yacobi *et al.*, 2024] Amitai Yacobi, Ofir Lindenbaum, and Uri Shaham. Specrage: Robust and generalizable multi-view spectral representation learning. *arXiv preprint arXiv:2411.02138*, 2024.

[Yadav *et al.*, 2021] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[Yang and Wang, 2018] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big data mining and analytics*, 1(2):83–107, 2018.

[Yue *et al.*, 2019] Zongsheng Yue, Hongwei Yong, Deyu Meng, Qian Zhao, Yee Leung, and Lei Zhang. Robust multiview subspace learning with nonindependently and nonidentically distributed complex noise. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1070–1083, 2019.

[Zhang *et al.*, 2018] Lei Zhang, Yao Zhao, Zhenfeng Zhu, Dinggang Shen, and Shuiwang Ji. Multi-view missing data completion. *IEEE Transactions on Knowledge and Data Engineering*, 30(7):1296–1309, 2018.

[Zhang *et al.*, 2020] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415, 2020.

[Zheng *et al.*, 2023] Qinghai Zheng, Jihua Zhu, Zhongyu Li, Zhiqiang Tian, and Chen Li. Comprehensive multi-view representation learning. *Information Fusion*, 89:198–209, 2023.