2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–17

# Variational Selective Autoencoder

**Yu Gong**[*†] GONGYUG@SFU.CA
**Hossein Hajimirsadeghi**[†] HOSSEIN.HAJIMIRSADEGHI@BOREALISAI.COM
**Jiawei He**[*†] JHA203@SFU.CA
**Megha Nawhal**[*†] MNAWHAL@SFU.CA
**Thibaut Durand**[†] THIBAUT.P.DURAND@BOREALISAI.COM
**Greg Mori**[*†] MORI@CS.SFU.CA

## Abstract

Despite promising progress on unimodal data imputation (e.g. image inpainting), models for multimodal data imputation are far from satisfactory. In this work, we propose variational selective autoencoder (VSAE) for this task. Learning only from partially-observed data, VSAE can model the joint distribution of observed/unobserved modalities and the imputation mask, resulting in a unified model for various down-stream tasks including data generation and imputation. Evaluation on synthetic high-dimensional and challenging low-dimensional multimodal datasets shows improvement over the state-of-the-art imputation models.

## 1. Introduction

Modern deep learning techniques rely heavily on extracting information from large scale datasets of clean and complete training data, such as labeled data or images with all pixels. Practically these data is costly due to the limited resources or privacy concerns. Having a model that learns and extracts information from partially-observed data will largely increase the application spectrum of deep learning models and provide benefit to down-stream tasks, e.g. data imputation, which has been an active research area. Despite promising progress, there are still challenges in learning effective imputation models: 1) Some prior works focus on learning from fully-observed data and then performing imputation on partially-observed data (Suzuki et al., 2016; Ivanov et al., 2019); 2) They usually have strong assumptions on missingness mechanism (see Appendix A.1) such as data is missing completely at random (MCAR) (Yoon et al., 2018); 3) Some other works explore only unimodal imputation such as image in-painting for high-dimensional data (Ivanov et al., 2019; Mattei and Frellsen, 2019).

Modeling any combination of data modalities has not been well-established yet. This can limit the potential of such models since raw data in real-life is usually acquired in a multimodal manner (Ngiam et al., 2011). A class of prior works focus on learning the conditional likelihood of the modalities (Sohn et al., 2015; Pandey and Dukkipati, 2017). However, they require complete data during training and cannot handle arbitrary conditioning. In practice, one or more of the modalities maybe be missing, leading to a challenging multimodal data imputation task. For more on related works, see Appendix A.2.

[*] Simon Fraser University
[†] Borealis AI

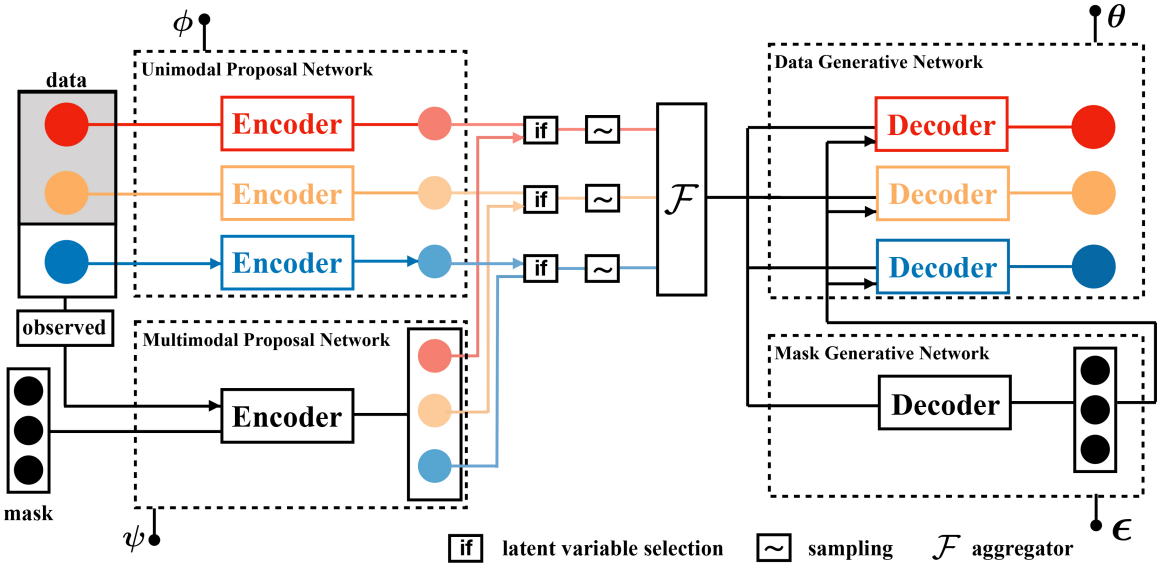


Figure 1: **Overall architecture.** Blue is observed, red/yellow are unobserved (shaded). The unimodal/multimodal proposal networks are employed by selection indicated by the arrows. Standard normal prior is ignored for simplicity. $\phi, \psi, \theta$ and $\epsilon$ are the parameters of each modules. All components are trained jointly.

We propose Variational Selective Autoencoder (VSAE) for multimodal data generation and imputation. It can model the joint distribution of data and mask and avoid the limited assumptions such as MCAR. VSAE is optimized efficiently with a single variational objective. The contributions are summarized as: (1) A novel variational framework to learn from partially-observed multimodal data; (2) VSAE can learn the joint distribution of observed/unobserved modalities and the mask, resulting in a unified model for various down-stream tasks including data generation/imputation with relaxed assumptions on missingness mechanism; (3) Evaluation on both synthetic high-dimensional and challenging low-dimensional multimodal datasets shows improvement over the state-of-the-art data imputation models.

## 2. Method

**Problem Statement.** Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_M]$ be the complete data with $M$ modalities. The size of each $\mathbf{x}_i$ may vary. A binary mask variable $\mathbf{m} \in \{0, 1\}^M$: $m_i = 1$ indicates $\mathbf{x}_i$ is observed and $m_i = 0$ indicates $\mathbf{x}_i$ is unobserved. The set of *observed modalities* $\mathbb{O} = \{i | m_i = 1\}$ and *unobserved modalities* $\mathbb{U} = \{i | m_i = 0\}$ are complementary. Accordingly, we denote the representation of observed and unobserved modalities with $\mathbf{x_o} = [\mathbf{x}_i | m_i = 1]$ and $\mathbf{x_u} = [\mathbf{x}_i | m_i = 0]$. Assuming $\mathbf{x}$ and $\mathbf{m}$ are dependent, we aim to model the joint distribution $p(\mathbf{x}, \mathbf{m})$. As a result, VSAE can be used for both imputation and generation.

**Proposed Model.** The high-level overview of VSAE (see Figure 1) is that the multimodal data is encoded to a latent space factorized w.r.t. the modalities. The latent variable of each modality is selectively chosen between a unimodal encoder (if the modality is observed) or a

multimodal encoder (if the modality is unobserved). Next all the modalities and mask are reconstructed by decoding the aggregated latent codes. Mathematically, we aim to model the joint distribution of the data $\mathbf{x} = [\mathbf{x_o}, \mathbf{x_u}]$ and mask $\mathbf{m}$. Following VAE formulation (see Appendix A.3), we derive the ELBO for $\log p(\mathbf{x}, \mathbf{m})$ with approximate posterior $q(\mathbf{z}|\mathbf{x}, \mathbf{m})$:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}) &= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log p_{\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}|\mathbf{z})] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}, \mathbf{m})||p(\mathbf{z})] \\ &= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m}, \mathbf{z}) + \log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}, \mathbf{m}) - \log p(\mathbf{z})].\end{aligned} \tag{1}$$

***Decoder*** The probability distribution factorizes over modalities assuming that reconstructions are conditionally independent given complete latent variables of all modalities:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m}, \mathbf{z}) = \log p_{\boldsymbol{\theta}}(\mathbf{x_o}, \mathbf{x_u}|\mathbf{m}, \mathbf{z}) = \sum_{i\in\mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z}) + \sum_{j\in\mathbb{U}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z}) \tag{2}$$

***Selective proposal distribution for encoders*** Following Tsai et al. (2019), we assume latent variables factorizes w.r.t the modalities $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M]$. Thus, $\log p(\mathbf{z}) = \sum_{i=1}^{M} \log p(\mathbf{z}_i)$ and $\log q(\mathbf{z}|\mathbf{x}, \mathbf{m}) = \sum_{i=1}^{M} \log q(\mathbf{z}_i|\mathbf{x}, \mathbf{m})$ . Given this, we define the proposal distribution for each modality as

$$q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) = \begin{cases} q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) & \text{if } m_i = 1 \\ q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m}) & \text{if } m_i = 0 \end{cases} \tag{3}$$

This is based on the intuitive assumption that the latent space of each modality is independent of the others given its data is observed. If the modality is missing, its latent variable is selectively inferred from other observed modalities. For partially-observed setting, $\mathbf{x_u}$ is unavailable even during training. Thus, we define the objective function for training by taking expectation of the ELBO in Equation (1) over $\mathbf{x_u}$. Only one term in Equation (2) is dependent on $\mathbf{x_u}$, so the final objective is derived as

$$\begin{aligned}\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o}, \mathbf{m}) &= \mathbb{E}_{\mathbf{x_u}}[\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})] = \mathbb{E}_{\mathbf{z}}\left[\sum_{i\in\mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m}, \mathbf{z}) + \sum_{j\in\mathbb{U}} \mathbb{E}_{\mathbf{x}_j}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})\right] \\ &+ \mathbb{E}_{\mathbf{z}}\left[\log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z})\right] - \sum_{i=1}^{M} \mathbb{E}_{\mathbf{z}_i}[\log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) - \log p(\mathbf{z}_i)] \quad \text{where } \mathbf{z}_i \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m})\end{aligned} \tag{4}$$

We approximate $\mathbb{E}_{\mathbf{x}_j}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})], j \in \mathbb{U}$ by sampling $\mathbf{x}_j$ from the prior network (standard normal) and passing through the decoder. Our experiments show even a single sample is sufficient to learn the model effectively. In fact, the prior network can be used as a self-supervision mechanism to find the most likely samples which dominate the other samples when taking the expectation. See Appendix B for more details.

## 3. Experiment

We evaluate our model on high-dimensional multi-modal data and low-dimensional tabular data in comparison with state-of-the-art latent variable models. To test the robustness of our model, we evaluate our model under various challenging missingness mechanisms.

| | Categorical(PFC) | | Numerical(NRMSE) | | Bimodal(MSE) | |
|---|---|---|---|---|---|---|
| | Phishing | Mushroom | Yeast | Glass | Synthetic1 | Synthetic2 |
| AE | $0.348 \pm 0.002$ | $0.56 \pm 0.01$ | $0.74 \pm 0.04$ | $1.65 \pm 0.05$ | $0.215 \pm \Delta$ | $0.234 \pm \Delta$ |
| VAE | $0.293 \pm 0.003$ | $0.47 \pm 0.02$ | $0.46 \pm \Delta$ | $1.41 \pm 0.01$ | $0.140 \pm \Delta$ | $0.127 \pm \Delta$ |
| CVAE w/ mask | $0.241 \pm 0.003$ | $0.45 \pm 0.01$ | $0.45 \pm \Delta$ | $1.50 \pm \Delta$ | $0.141 \pm \Delta$ | $0.125 \pm \Delta$ |
| MVAE (Wu and Goodman, 2018) | $0.308 \pm 0.015$ | $0.59 \pm 0.02$ | $0.44 \pm 0.02$ | $1.57 \pm 0.04$ | $0.156 \pm \Delta$ | $0.131 \pm \Delta$ |
| VSAE | $\mathbf{0.237 \pm 0.001}$ | $\mathbf{0.40 \pm 0.01}$ | $\mathbf{0.41 \pm 0.01}$ | $\mathbf{1.31 \pm 0.02}$ | $\mathbf{0.138 \pm \Delta}$ | $\mathbf{0.120 \pm \Delta}$ |
| CVAE w/ data | $0.301 \pm 0.005$ | $0.49 \pm 0.03$ | $0.45 \pm \Delta$ | $1.38 \pm 0.05$ | $0.134 \pm \Delta$ | $0.127 \pm \Delta$ |
| VAEAC(Ivanov et al., 2019) | $0.240 \pm 0.006$ | $0.40 \pm 0.01$ | $0.45 \pm \Delta$ | $1.43 \pm 0.03$ | $0.134 \pm \Delta$ | $0.121 \pm \Delta$ |

Table 1: **Data Imputation.** Missing ratio is 0.5. For all lower is better. Last two rows are trained with fully-observed data. We show mean/std over 3 independent runs. $\Delta \leq 0.001$.

### 3.1. Data Imputation

**Low-dimensional tabular data.** We choose UCI repository datasets (contains both numerical and categorical data, training/test split as 4/1 and 20% of training set for validation). We randomly sample from independent Bernoulli distributions with pre-defined missing ratio to generate masks that are fixed during training and test. In Table 1, we observe that VSAE trained with partially-observed data outperforms other baselines, even those models trained with fully-observed data on some datasets. We argue this is due to two potential reasons: (1) the mask provides a natural way of dropout on the data space, thereby, helping the model to generalize; (2) If the data is noisy or has outliers, which is common in low-dimensional data, learning from partially-observed data can improve the results by ignoring these data. Figure 2 indicates VSAE are more robust to missing ratio.

**High-dimensional multimodal data.** We synthesize two bimodal datasets using MNIST and SVHN datasets: **Synthetic1** contains randomly paired two digits in MNIST as [0, 9], [1, 8], [2, 7], [3, 6], [4, 5]; **Synthetic2** contains pairs of one digit in MNIST with a random same digit in SVHN. VSAE has better performance with lower variance (see Table 1). Experiments also indicate that VSAE is robust under different missing ratios, whereas other baselines are sensitive to the missing ratio, which is consistent as UCI experiments. We believe this is because of the underlying mechanism of proper proposal distribution selection and prior sharing. The separation of unimodal/multimodal encoders helps the model to attend to the observed data, while baselines only have single proposal distribution inferred from the whole input. Thus, VSAE can easily ignore unobserved noisy modalities and attends on observable useful modalities, but baselines rely on neural networks to extract useful information from the whole data (which is dominated by missing information in case of high missing ratio).

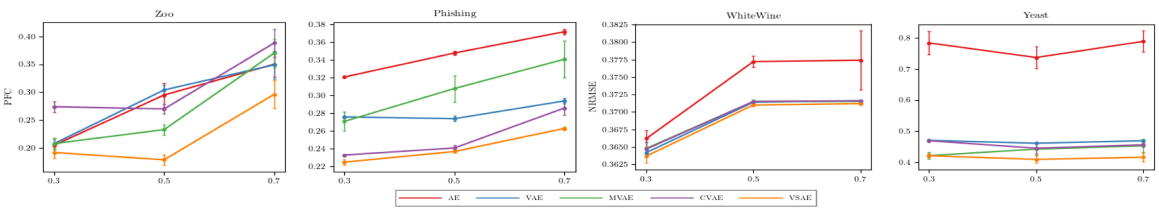


Figure 2: **Feature Imputations** in different missing ratios, over 3 independent runs.

Figure 3: **Data Generation.** Generated w/o conditional information. As shown, the correspondence between modalities (pre-defined pairs) are preserved.

### 3.2. Data and Mask Generation

After training, we sample from prior to generate the data and mask. In UCI experiments, We calculate the proportion of 0 in generated mask vectors over 100 samples and average on all experiments, we get $0.3123 \pm 0.026$, $0.4964 \pm 0.005$, $0.6927 \pm 0.013$ for missing ratio of 0.3, 0.5, 0.7. It indicates VSAE can learn mask distribution. We observe that conditions on the reconstructed mask in the data decoders improve the performance. We believe this is because the mask vector can inform the data decoder about the missingness in data space since the latent space is shared by all modalities thereby allowing it to generate data from the selective proposal distribution. Figure 3 qualitatively shows VSAE can perform good data generation in bimodal experiments.

## References

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations*, 2016.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hong-Min Chu, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Deep generative models for weakly-supervised multi-label classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Lovedeep Gondara and Ke Wang. Multiple imputation using deep denoising autoencoders. *arXiv preprint arXiv:1705.02737*, 2017.

Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Machine Learning*, 2019.

Vikas Jain, Nirbhay Modhe, and Piyush Rai. Scalable generative models for multi-label learning with missing labels. In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-80254-9.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 4234–4243, 2019.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423, 2019.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM, 2011.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, ICML'11, pages 689–696, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 308–315. IEEE, 2017.

Casper Kaae Sø nderby, Tapani Raiko, Lars Maalø e, Søren Kaae Sø nderby, and Ole Winther. Ladder variational autoencoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3738–3746. Curran Associates, Inc., 2016.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in neural information processing systems*, pages 2141–2149, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28*, pages 3483–3491. 2015.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 2018.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Yao-Chong Lim, Paul Pu Liang, and Louis-Philippe Morency. Variational auto-decoder: Neural generative modeling from partial data. *arXiv preprint arXiv:1903.00840*, 2019.

## Appendix A. Backgroud

### A.1. Imputing data from missing information

**Imputation Process and Missingness Mechanisms.** Following (Little and Rubin, 1986), the imputation process is to learn a generative distribution for unobserved missing data. To be consistent with notations in Section 2, let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]$ be the complete data of all modalities where $\mathbf{x}_i$ denote the feature representation for the $i$-th modality. We also define $\mathbf{m} \in \{0, 1\}^M$ as the binary mask vector, where $m_i = 1$ indicates if the $i$-th modality is observed, and $m_i = 0$ indicates if it is unobserved:

$$\begin{aligned} \mathbf{x} &\sim p_{\text{data}}(\mathbf{x}), \\ \mathbf{m} &\sim p(\mathbf{m}|\mathbf{x}). \end{aligned} \tag{5}$$

Given this, the observed data $\mathbf{x_o}$ and unobserved data $\mathbf{x_u}$ are represented accordingly:

$$\begin{aligned} \mathbf{x_o} &= [\mathbf{x}_i | m_i = 1], \\ \mathbf{x_u} &= [\mathbf{x}_i | m_i = 0]. \end{aligned} \tag{6}$$

In the standard maximum likelihood setting, the unknown parameters are estimated by maximizing the following marginal likelihood, integrating over the unknown missing data values:

$$p(\mathbf{x_o}, \mathbf{m}) = \int p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) d\mathbf{x_u} = \int p(\mathbf{x_o}, \mathbf{x_u}) p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) d\mathbf{x_u} \tag{7}$$

Little and Rubin (1986) characterizes the missingness mechanism $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u})$ in terms of independence relations between the complete data $\mathbf{x} = [\mathbf{x_o}, \mathbf{x_u}]$ and the mask $\mathbf{m}$:

- Missing completely at random (MCAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m})$,
- Missing at random (MAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m}|\mathbf{x_o})$,
- Not missing at random (NMAR): $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) = p(\mathbf{m}|\mathbf{x_u})$ or $p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u})$.

Most previous data imputation methods works under MCAR or MAR assumptions since $p(\mathbf{x_o}, \mathbf{m})$ can be factorized into $p(\mathbf{x_o})p(\mathbf{m}|\mathbf{x_o})$ or $p(\mathbf{x_o})p(\mathbf{m})$. With such decoupling, we do not need missing information to marginalize the likelihood, and it provides a simple but approximate framework to learn from partially-observed data.

**Data Imputation.** Classical imputation methods such as MICE (Buuren and Groothuis-Oudshoorn, 2010) and MissForest (Stekhoven and Bühlmann, 2011) learn discriminative models to impute missing features from observed ones. With recent advances in deep learning, several deep imputation models have been proposed based on autoencoders (**?**Gondara and Wang, 2017; Ivanov et al., 2019), generative adversarial nets (GANs) (Yoon et al., 2018; Li et al., 2019), and autoregressive models (**?**). GAN-based imputation method GAIN proposed by Yoon et al. (2018) assumes that data is missing completely at random. Moreover, this method does not scale to high-dimensional multimodal data. Several VAE based data imputation methods (Ivanov et al., 2019; Nazabal et al., 2018; Mattei and Frellsen, 2019) have been proposed in recent years. Ivanov et al. (2019) formulated variational autoencoders with arbitrary conditioning (VAEAC) for data imputation which allows generation of missing

data conditioned on any combination of observed data. This algorithm needs complete data during training cannot learn from partially-observed data only. Nazabal et al. (2018) and Mattei and Frellsen (2019) modified VAE formulation to model the likelihood of the observed data only. However, they require training of a separate generative network for each dimension thereby increasing computational requirements. In contrast, our method aims to model joint distribution of observed and unobserved data along with the missingness pattern (imputation mask). This enables our model to perform both data generation and imputation even under relaxed assumptions on missingness mechanism (see Appendix A.1).

### A.2. Learning from Multimodal Data.

A class of prior works such as conditional VAE (Sohn et al., 2015) and conditional multimodal VAE (Pandey and Dukkipati, 2017) focus on learning the conditional likelihood of the modalities. However, these models requires complete data during training and cannot handle arbitrary conditioning. Alternatively, several generative models aim to model joint distribution of all modalities (Ngiam et al., 2011; **?**; Sohn et al., 2014; Suzuki et al., 2016). However, multimodal VAE based methods such as joint multimodal VAE (Suzuki et al., 2016) and multimodal factorization model (MFM) (Tsai et al., 2019) require complete data during training. On the other hand, Wu and Goodman (2018) proposed another multimodal VAE (namely MVAE) can be trained with incomplete data. This model leverages a shared latent space for all modalities and obtains an approximate joint posterior for the shared space assuming each modalities to be factorized. However, if training data is complete, this model cannot learn the individual inference networks and consequently does not learn to handle missing data during test. Building over multimodal VAE approaches, our model aims to address the shortcomings above within a flexible framework. In particular, our model can learn multimodal representations from partially observed training data and perform data imputation from arbitrary subset of modalities during test. By employing a factorized multimodal representations in the latent space it resembles disentangled models which can train factors specialized for learning from different parts of data (Tsai et al., 2019).

### A.3. Variational Autoencoder

Variational Autoencoder (VAE) (Kingma and Welling, 2013) is a probabilistic generative model, where data is constructed from a latent variable $\mathbf{z}$ with a prior distribution $p(\mathbf{z})$. It is composed of an inference network and a generation network to encode and decode data. To model the likelihood of data, the true intractable posterior $p(\mathbf{z}|\mathbf{x})$ is approximated by a proposal distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, and the whole model is trained until ideally the decoded reconstructions from the latent codes sampled from the approximate posterior match the training data. In the generation module, $p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}|\mathbf{z})$, a decoder realized by a deep neural network parameterized by $\boldsymbol{\theta}$, maps a latent variable $\mathbf{z}$ to the reconstruction $\tilde{\mathbf{x}}$ of observation $\mathbf{x}$. In the inference module, an encoder parameterized by $\boldsymbol{\phi}$ produces the sufficient statistics of the approximation posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (a known density family where sampling can be readily done). In vanilla VAE setting, by simplifying approximate posterior as a parameterized diagonal normal distribution and prior as a standard diagonal normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the training criterion is to maximize the following evidence lower bound (ELBO) w.r.t. $\boldsymbol{\theta}$

and $\boldsymbol{\phi}$.

$$\log p(\mathbf{x}) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{8}$$

where $D_{\mathrm{KL}}$ denotes the Kullback-Leibler (KL) divergence. Usually the prior $p(\mathbf{z})$ and the approximate $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ are chosen to be in simple form, such as a Gaussian distribution with diagonal covariance, which allows for an analytic calculation of the KL divergence. While VAE approximates $p(\mathbf{x})$, conditional VAE (Sohn et al., 2015) approximates the conditional distribution $p(\mathbf{x}|\mathbf{y})$. By simply introducing a conditional input, CVAE is trained to maximize the following ELBO:

$$\log p(\mathbf{x}|\mathbf{y}) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{x},\mathbf{y}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\mathbf{y})] - D_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y})||p_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{y})] \tag{9}$$

## Appendix B. Module Details

### B.1. Architecture

We construct each module of our model using neural networks and optimize the parameters via backpropagation techniques. Following the terms in standard VAE, our model is composed of encoders and decoders. The architecture is shown in Figure 1 with different modalities denoted by different colors. The data space of unobserved modalities is shaded to differentiate from observed modalities. The whole architecture can be viewed as an integration of two auto-encoding structures: the top-branch data-wise encoders/decoders and the bottom-branch mask-wise encoders/decoder. The selective proposal distribution chooses between the unimodal and multimodal encoders if the data is observed or not. The outputs of all encoders are aggregated and a common latent space is shared among all decoders. In the rest of this section we explain different modules in the proposed model. For more details about architecture and implementation see Appendix B.

**Selective Factorized Encoders** Standard proposal distribution of VAEs depends on the whole data and can not handle incomplete input when the data is partially-observed. To overcome this, we introduce our selective proposal distribution, which is factorized with respect to the modalities. As defined in Equation (3), $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$, named as the *unimodal proposal distribution*, is inferred only from each observed individual modality of data. However, if the modality is unobserved, the *multimodal proposal distribution* $q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m})$ is used to infer corresponding latent variables from other observed modalities and mask. Hence, the learned model can impute the missing information by combining unimodal proposal distribution of observed modalities and multimodal proposal distribution of the unobserved modalities. The condition on the mask could make the model aware of the missing pattern and could help the model to attend to observed modalities.

For each modality $\mathbf{x}_i$, we have a separate encoder to infer its unimodal proposal distribution parameterized by $\boldsymbol{\phi}$. For the multimodal proposal distribution, however, we use a single encoder parameterized by $\boldsymbol{\psi}$. This encoder outputs the latent codes for all modalities, and we simply obtain the latent variable for each modality by slicing the output vector to $M$ sequential vectors. We simply model all the proposal distributions as normal distributions by setting the outputs of all encoders as mean and variance of a normal distribution. For the unimodal proposal distributions, we have $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_i), \boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathbf{x}_i))$, where $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ are deterministic neural networks parameterized by $\boldsymbol{\phi}$ that output the mean and covariance respectively. Similarly, the multimodal proposal distribution $q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m}) = \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{\boldsymbol{\psi}}(\mathbf{x_o}, \mathbf{m}), \boldsymbol{\Sigma}_{\boldsymbol{\psi}}(\mathbf{x_o}, \mathbf{m}))$ can be modeled by a neural network with $\mathbf{x_o}$ and $\mathbf{m}$ as the inputs. The reparameterization in standard VAE is used for end-to-end training.

**Decoding through Latent Variable Aggregator $\mathcal{F}$** After selecting and sampling from proper proposal distributions for all modalities, the variational latent codes can be fed to the downstream decoders even when the observation is incomplete. To do this, the information from different modalities interact by aggregating their stochastic latent codes before going

through the decoders:

$$p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z}) = p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathcal{F}(\mathbf{z})),$$
$$p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z},\mathbf{m}) = p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathcal{F}(\mathbf{z}),\mathbf{m})), \tag{10}$$

Here we simply choose the aggregator $\mathcal{F}(\cdot) = \text{concat}(\cdot)$, i.e., concatenating the latent codes as one single vector. One may also use other aggregation functions such as max/mean pooling or matrix fusion (Veit et al., 2018) to combine latent codes from all modalities. The decoders take the shared aggregated variational latent codes as input to generate data and mask.

**Mask Vector Encoding and Decoding** The mask variable $\mathbf{m}$ is encoded into the latent space through the multimodal proposal network. The latent space is shared by the mask and data decoders. The mask decoder is an MLP parameterized by $\boldsymbol{\epsilon}$ in our implementation. It maps the aggregated latent codes from the selective proposal distributions to a reconstruction of the $M$-dimensional binary mask vector. We assume each dimension of the mask variable is an independent Bernoulli distribution.

**Training** With reparameterization trick (Kingma and Welling, 2013), we can jointly optimize the objective derived in Equation Equation (4) with respect to these parameters defined above on training set:

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}} \mathbb{E}_{\mathbf{x_o},\mathbf{m}}[\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}}(\mathbf{x_o},\mathbf{m})] \tag{11}$$

Since Equation (11) only requires the mask and observed data during training, this modified ELBO $\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\epsilon}}(\mathbf{x_o},\mathbf{m})$ can be optimized without the presence of unobserved modalities. The KL-divergence term is calculated analytically for each factorized term. The conditional log-likelihood term is computed by negating reconstruction loss function. (See Section 3 and Appendix B.3.)

**Inference** The learned model can be used for both data imputation and generation. For imputation, the observed modalities $\mathbf{x_o}$ and mask $\mathbf{m}$ are fed through the encoders to infer the selective proposal distributions. Then the sampled latent codes are decoded to estimate the unobserved modalities $\mathbf{x_u}$. All the modules in Figure 1 are used for imputation. For generation, since no data is available at all, the latent codes are sampled from the prior and go through the decoders to generate the data and the mask. In this way, only modules after the aggregator are used without any inference modules.

### B.2. Implementation

In all models, all the layers are modeled by MLP without any skip connections or resnet modules. Basically, the unimodal encoders take single modality data vector as input to infer the unimodal proposal distribution; the multimodal encoders take the observed data vectors and mask vector as as input to infer the multimodal proposal distributions. The input vector to multimodal encoders should have same length for the neural network. Here we just concatenate all modality vectors and replace the unobserved modality vectors with some noise. In UCI repository experiment, we replace the unobserved modality vectors as

standard normal noise. In Bimodal experiment, we simply replace the pixels of unobserved modality as zero. Note that all the baselines has encoders/decoders with same or larger number of parameters than our method. We implement our model using PyTorch.

**Unimodal Encoders** In UCI repository experiment, the unimodal encoders for numerical data are modeled by 3-layer 64-dim MLPs and the unimodal encoders for categorical data are modeled by 3-layer 64-dim MLPs, all followed by Batch Normalization and Leaky ReLU nonlinear activations. In MNIST+MNIST bimodal experiment, the unimodal encoders are modeled by 3-layer 128-dim MLPs followed by Leaky ReLU nonlinear activations; In MNIST+SVHN bimodal experiment, the unimodal encoders are modeled by 3-layer 512-dim MLPs followed by Leaky ReLU nonlinear activations. We set the latent dimension as 20-dim for every modality in UCI repository experiments and 256-dim for every modality in Bimodal experiments.
UCI data unimodal encoder: Linear(1, 64)→ BatchNorm1d(64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 20);
MNIST+MNIST synthetic unimodal encoder: Linear(data-dimension, 128)→ LeakyReLU→ Linear(128,128)→ LeakyReLU→ Linear(128, 128)→ LeakyReLU→ Linear(128, 256);
MNIST+SVHN synthetic unimodal encoder: Linear(data-dimension, 512)→ LeakyReLU→ Linear(512,512)→ LeakyReLU→ Linear(512, 512)→ LeakyReLU→ Linear(512, 256);

**Multimodal Encoders** In general, any model capable of multimodal fusion (Zadeh et al., 2017; Morency et al., 2011) can be used here to map the observed data $\mathbf{x_o}$ and the mask $\mathbf{m}$ to the latent variables $\mathbf{z}$. However, in this paper we simply use an architecture similar to unimodal encoders. The difference is that the input to unimodal encoders are lower dimensional vectors of an individual modalities. But, the input to the multimodal encoders is the complete data vector with unobserved modalities replaced with noise or zeros. As the input to the multimodal encoders is the same for all modalities (i.e., $q(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m})\ \forall i$), we model the multimodal encoders as one single encoder to take advantage of the parallel matrix calculation speed. Thus the multimodal encoder for every experiment has the same structure as its unmidal encoder but with full-dimensional input.

**Aggregator** In our models, we simply use vector concatenation as the way of aggregating.

**Mask Decoder** UCI mask decoder: Linear(20*data-dimension, 64)→ BatchNorm1d(64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, 64)→ LeakyReLU→ Linear(64, mask-dimension)→Sigmoid;
MNIST+MNIST synthetic mask decoder: Linear(512, 16)→ BatchNorm1d(16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16, 16)→ LeakyReLU→ Linear(16, 2)→Sigmoid;
MNIST+SVHN synthetic mask encoder: Linear(512, 16)→ BatchNorm1d(16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16,16)→ LeakyReLU→ Linear(16,2)→Sigmoid;

**Data Decoder** As the output is factorized over modalities and for every decoder the input is shared as the latent codes sampled from the selective proposal distribution. We implement all the decoders of the data modalities as one single decoder for parallel speed.
UCI data decoder: Linear(20*data-dimension, 128)→ BatchNorm1d(128)→ LeakyReLU→ Linear(128)→ Linear(128, 128)→ Linear(128, data-dimension);
MNIST+MNIST synthetic data decoder: Linear(512, 128)→ BatchNorm1d(128)→ LeakyReLU→

Linear(128,128)→ Linear(128, 128)→ Linear(128, 784)→Sigmoid;
MNIST+SVHN synthetic mask encoder: Linear(512, 512)→ BatchNorm1d(512)→ LeakyReLU→ Linear(512,512)→ Linear(512,512)→ Linear(512,784/3072)→Sigmoid;

### B.3. Training

We use Adam optimizer for all models. For UCI numerical experiment, learning rate is 1e-3 and use validation set to find a best model in 1000 epochs. For UCI categorical experiment, learning rate is 1e-2 and use validation set to find a best model in 1000 epochs. For bimodal experiments, learning rate is 1e-4 and use validation set to find a best model in 1000 epochs. All modules in our models are trained jointly.
In our model, we calculate the conditional log-likelihood of unobserved modality by generating corresponding modalities from prior. We initially train the model for some (empirically we choose 20) epochs without calculating the conditional log-likelihood of $\mathbf{x_u}$. And then first feed the partially-observed data to the model and generate the unobserved modality $\tilde{\mathbf{x}}_\mathbf{u}$ without calculating any loss; then feed the same batch for another pass, calculate the conditional log-likelihood using real $\mathbf{x_o}$ and generated $\mathbf{x_u}$ as ground truth.

## Appendix C. Additional Results

### C.1. UCI repository Datasets

| | Phishing | Zoo | Mushroom |
|---|---|---|---|
| AE | $0.348 \pm 0.002$ | $0.295 \pm 0.022$ | $0.556 \pm 0.009$ |
| VAE | $0.293 \pm 0.003$ | $0.304 \pm 0.009$ | $0.470 \pm 0.017$ |
| CVAE w/ mask | $0.241 \pm 0.003$ | $0.270 \pm 0.023$ | $0.445 \pm 0.004$ |
| MVAE | $0.308 \pm 0.015$ | $0.233 \pm 0.013$ | $0.586 \pm 0.019$ |
| VSAE | $\mathbf{0.237 \pm 0.001}$ | $\mathbf{0.213 \pm 0.004}$ | $\mathbf{0.396 \pm 0.008}$ |
| CVAE w/ data | $0.301 \pm 0.005$ | $0.323 \pm 0.032$ | $0.485 \pm 0.034$ |
| VAEAC | $0.240 \pm 0.006$ | $0.168 \pm 0.006$ | $0.403 \pm 0.006$ |

Table 2: **Imputation on Categorical datasets**. Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by PFC, lower is better.

| | Yeast | White Wine | Glass |
|---|---|---|---|
| AE | $0.737 \pm 0.036$ | $0.3772 \pm 0.0008$ | $1.651 \pm 0.049$ |
| VAE | $0.461 \pm 0.001$ | $0.3714 \pm 0.0001$ | $1.409 \pm 0.011$ |
| CVAE w/ mask | $0.449 \pm 0.001$ | $0.3716 \pm 0.0001$ | $1.498 \pm 0.0013$ |
| MVAE | $0.442 \pm 0.018$ | $\mathbf{0.3722 \pm 0.0009}$ | $1.572 \pm 0.035$ |
| VSAE | $\mathbf{0.409 \pm 0.012}$ | $\mathbf{0.3711 \pm 0.0002}$ | $\mathbf{1.312 \pm 0.021}$ |
| CVAE w/ data | $0.449 \pm 0.0001$ | $0.3567 \pm 0.0016$ | $1.380 \pm 0.045$ |
| VAEAC | $0.447 \pm 0.0016$ | $0.3647 \pm 0.0039$ | $1.432 \pm 0.027$ |

Table 3: **Imputation on Numerical datasets**. Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by NRMSE, lower is better.

### C.2. MNIST+MNIST Bimodal dataset

| | 0.3 | 0.5 | 0.7 |
|---|---|---|---|
| AE | $0.2124 \pm 0.0012$ | $0.2147 \pm 0.0008$ | $0.2180 \pm 0.0008$ |
| VAE | $0.1396 \pm 0.0002$ | $0.1416 \pm 0.0001$ | $0.1435 \pm 0.0006$ |
| CVAE w/ mask | $0.1393 \pm 0.0002$ | $0.1412 \pm 0.0006$ | $0.1425 \pm 0.0012$ |
| MVAE | $0.1547 \pm 0.0012$ | $0.1562 \pm 0.0003$ | $0.1579 \pm 0.0006$ |
| VSAE | $\mathbf{0.1371 \pm 0.0001}$ | $\mathbf{0.1376 \pm 0.0002}$ | $\mathbf{0.1379 \pm 0.0001}$ |
| CVAE w/ data | $0.1336 \pm 0.0003$ | $0.1340 \pm 0.0003$ | $0.1343 \pm 0.0002$ |
| VAEAC | $0.1333 \pm 0.0004$ | $0.1338 \pm 0.0003$ | $0.1344 \pm 0.0001$ |

Table 4: **Imputation on MNIST+MNIST.** Missing ratio is 0.3, 0.5 and 0.7. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.

Figure 4: **Imputation on MNIST+MNIST.** Top row visualizes observed modality, middle row unobserved modality, and bottom row shows the imputation of unobserved modality from VSAE.

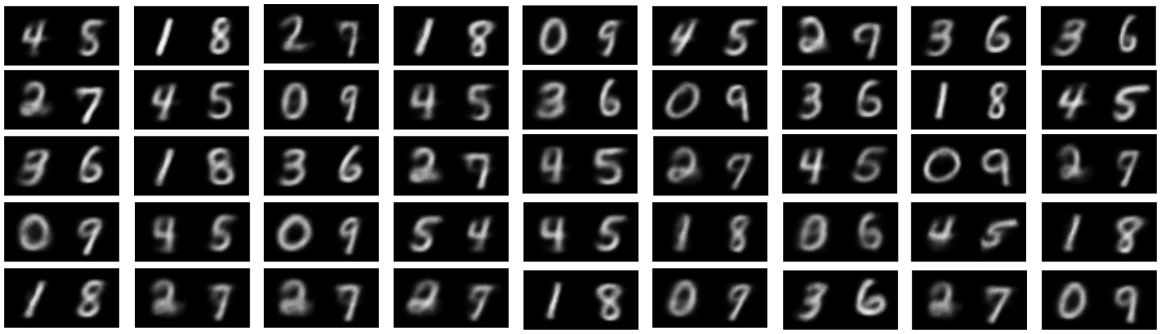

Figure 5: **Generation on MNIST+MNIST.** Generated Samples w/o conditional information. As shown, the correspondence between modalities (pre-defined pairs) are preserved while generation.

### C.3. MNIST+SVHN Bimodal dataset

| | MNIST-MSE/784 | SVHN-MSE/3072 | Bimodal Error |
|---|---|---|---|
| AE | $0.0867 \pm 0.0001$ | $0.1475 \pm 0.0006$ | $0.2342 \pm 0.0007$ |
| VAE | $0.0714 \pm 0.0001$ | $0.0559 \pm 0.0027$ | $0.1273 \pm 0.0003$ |
| CVAE w/ mask | $0.0692 \pm 0.0001$ | $0.0558 \pm 0.0003$ | $0.1251 \pm 0.0005$ |
| MVAE | $0.0707 \pm 0.0003$ | $0.602 \pm 0.0001$ | $0.1309 \pm 0.0005$ |
| VSAE | $\mathbf{0.0682 \pm 0.0001}$ | $\mathbf{0.0516 \pm 0.0001}$ | $\mathbf{0.1198 \pm 0.0001}$ |
| CVAE w/ data | $0.0716 \pm 0.0002$ | $0.0550 \pm 0.0007$ | $0.1266 \pm 0.0005$ |
| VAEAC | $0.0682 \pm 0.0001$ | $0.0523 \pm 0.0001$ | $0.1206 \pm 0.0001$ |

Table 5: **Imputation on MNIST+SVHN.** Missing ratio is 0.5. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.

| | 0.3 | 0.5 | 0.7 |
|---|---|---|---|
| AE | $0.1941 \pm 0.0006$ | $0.2342 \pm 0.0007$ | $0.2678 \pm 0.0012$ |
| VAE | $0.1264 \pm 0.0001$ | $0.1273 \pm 0.0003$ | $0.1322 \pm 0.0005$ |
| CVAE w/ mask | $0.1255 \pm 0.0002$ | $0.1251 \pm 0.0005$ | $0.1295 \pm 0.0006$ |
| MVAE | $0.1275 \pm 0.0029$ | $0.1309 \pm 0.0005$ | $0.1313 \pm 0.0013$ |
| VSAE | $\mathbf{0.1217 \pm 0.0002}$ | $\mathbf{0.1198 \pm 0.0001}$ | $\mathbf{0.1202 \pm 0.0002}$ |
| CVAE w/ data | $0.1288 \pm 0.0011$ | $0.1266 \pm 0.0005$ | $0.1248 \pm 0.0003$ |
| VAEAC | $0.1218 \pm 0.0002$ | $0.1206 \pm 0.0001$ | $0.1211 \pm 0.0001$ |

Table 6: **Imputation on MNIST+SVHN.** Missing ratio is 0.3, 0.5 and 0.7. Last two rows are trained with fully-observed data. Evaluated by combined errors of two modalities, lower is better.