





# MultiVae: A Python package for Multimodal Variational Autoencoders on Partial Datasets.

Agathe Senellart <sup>1,2,3</sup>✉, Clément Chadebec<sup>1,2,3</sup>, and Stéphanie Allasonnière<sup>1,2,3</sup>

<sup>1</sup> Université de Paris-Cité <sup>2</sup> Inria <sup>3</sup> Inserm ✉ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

In recent years, multimodal machine learning has seen significant growth, especially in representation learning and data generation. Recently, Multimodal Variational Autoencoders (VAEs) have been attracting growing interest for both tasks, thanks to their versatility, scalability, and interpretability as latent variable models. They are particularly useful in *partially observed* settings, such as medical applications, where available datasets are often incomplete ([Antelmi et al., 2019](#); [Lawry Aguila et al., 2023](#)).

We introduce MultiVae, an open-source Python library offering unified implementations of multimodal VAEs. It is designed for easy and customizable use of these models on fully or partially observed data. It facilitates the development and benchmarking of new algorithms by including standard benchmark datasets, evaluation metrics and tools for monitoring and sharing models.

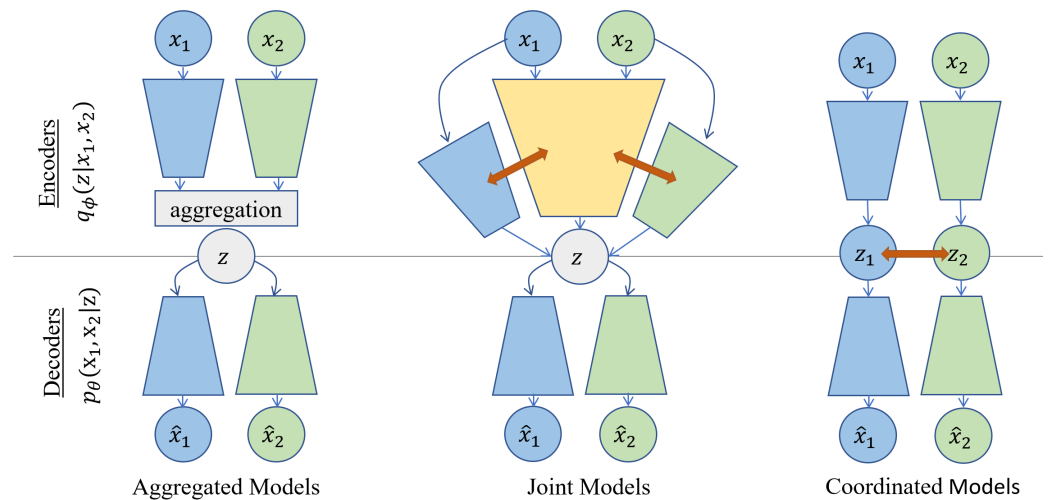
## Multimodal Variational Autoencoders

Multimodal VAEs aim to: (1) Learn a shared representation from multiple modalities; (2) Generate one missing modality from available ones.

These models learn a latent representation  $z$  of all modalities in a lower dimensional space and learn to *decode*  $z$  to generate each modality.

Let  $X = (x_1, x_2, \dots, x_M)$  contain  $M$  modalities. In the VAE setting, we define an *encoder* distribution  $q_\phi(z|X)$  projecting the observations to the latent space, and decoders distributions  $(p_\theta(x_i|z))_{1 \leq i \leq M}$  translating the latent code  $z$  back to observations. Those distributions are parameterized by neural networks that are trained to minimize an objective function derived from variational inference. See ([Kingma & Welling, 2014](#)) to learn more about the VAE framework and ([Suzuki & Matsuo, 2022](#)) for a survey on multimodal VAEs.

A key differentiator of multimodal VAEs relies in the choice of the encoder  $q_\phi(z|X)$ . They fall into three main categories: *Aggregated models* ([Shi et al., 2019](#); [Sutter et al., 2021](#); [Wu & Goodman, 2018](#)) use a mean or product operation to combine modalities, *Joint models* ([Senellart et al., 2023](#); [Suzuki et al., 2016](#); [Vedantam et al., 2018](#)) use a neural network taking all modalities as input, and *Coordinated models* ([Tian & Engel, 2019](#); [Wang et al., 2017](#)) use separate latent spaces with additional similarity constraints.



MultiVae unifies these approaches in a modular and extensible way.

Notably, aggregated models offer a natural way of *learning* on incomplete datasets: for an incomplete sample  $X$ , the encoding  $z$  and the objective function can be computed using only available modalities. MultiVae is the first library to provide implementations of these models with built-in support for missing data, using masks during loss computation.

## Data Augmentation

Another application of VAEs is Data Augmentation (DA): by sampling new latent codes  $z$  and decoding them, *fully synthetic multimodal* samples can be generated to augment a dataset. This approach has been successfully used with unimodal VAEs to augment datasets for data-intensive deep learning applications (Chadebec et al., 2023). However, it remains underexplored in the multimodal setting. MultiVae includes a `multivae.samplers` module with several sampling strategies to further explore the generative abilities of these models.

## Statement of Need

Despite the usefulness of multimodal VAEs, the lack of easy-to-use and verified implementations might hinder applicative research. MultiVae offers unified implementations, designed to be accessible even for non-specialists. We ensure reliability by reproducing key results from original papers whenever possible.

Related libraries contain implementations of Multimodal VAEs: the [Multimodal VAE Comparison Toolkit](#) (Sejnova et al., 2024), [Pixyz](#) (Masahiro Suzuki & Matsuo, 2023) and [multi-view-ae](#) (Aguila et al., 2023) that is most closely related to us and released while we were developing MultiVae.

We compare in a summarizing table below, the different features of each work. MultiVae differs and complements existing software packages in key ways: it supports **incomplete datasets**, which we consider essential for real-life applications, as well as **generative samplers**, **benchmark datasets** and **metrics** to facilitate research. It contains a large range of models with a great flexibility on parameters' choices and including all implementation details present in the original codes that improve performance.

63 **List of Models and Features**

64 We list models and features in each work. Symbol (✓\*) indicates that the implementation  
65 includes additional options unavailable in the others.

Models/ Features	Ours	(Aguila et al., 2023)	(Sejnova et al., 2024)	(Masahiro Suzuki & Matsuo, 2023)
JMVAE(Suzuki et al., 2016)	✓*	✓		✓
MVAE(Wu & Goodman, 2018)	✓*	✓	✓	✓
MMVAE(Shi et al., 2019)	✓*	✓	✓	
MoPoE(Sutter et al., 2021)	✓*	✓	✓	
DMVAE(Lee & Pavlovic, 2021)	✓	✓*	✓	
MVTCAE(Hwang et al., 2021)	✓	✓		
MMVAE+(Palumbo et al., 2023)	✓*	✓		
CMVAE(Palumbo et al., 2024)	✓			
Nexus(Vasco et al., 2022)	✓			
CVAE(Kingma & Welling, 2014)	✓			✓
MHVAE(Dorent et al., 2023)	✓			
TELBO(Vedantam et al., 2018)	✓			
JNF(Senellart et al., 2023)	✓			
CRMVAE(Suzuki & Matsuo, 2023)	✓			
MCVAE(Antelmi et al., 2019)		✓		
mAAE		✓		
DVCCA(Wang et al., 2017)		✓		
DCCAE(Wang et al., 2015)		✓		
mWAE		✓		
mmJSD(Sutter et al., 2020)		✓		
gPoE(Lawry Aguila et al., 2023)		✓		
Support of Incomplete datasets	✓			
GMM Sampler	✓			
MAF Sampler, IAF Sampler	✓			
Metrics: {Likelihood, Coherences, FIDs, Reconstruction, Clustering}	✓			
Benchmark Datasets	✓		✓	
Model sharing via Hugging Face	✓			

66 **Code Quality and Documentation**

67 MultiVae is available on [Github](#) and Pypi, with full documentation at (<https://multivae.readthedocs.io/>). The code is unit-tested with 94% coverage. We provide **tutorials** either as notebooks  
68 or scripts allowing users to get started easily. To further showcase how to use our library for  
69 research applications, we provide detailed *case studies* in the documentation.  
70

71 **Acknowledgements**

72 We are grateful to the authors of the initial implementations of the models included in MultiVae.  
73 This work benefited from state grant managed by the Agence Nationale de la Recherche under  
74 the France 2030 program, AN-23-IACL-0008. This research has been partly supported by the  
75 European Union under the (2023-2030) ERC Synergy Grant 101071601.

## References

- Aguila, A. L., Jayme, A., Montaña-Brown, N., Heuveline, V., & Altmann, A. (2023). Multi-view-AE: A python package for multi-view autoencoder models. *Journal of Open Source Software*, 8(85), 5093. <https://doi.org/10.21105/joss.05093>
- Antelmi, L., Ayache, N., Robert, P., & Lorenzi, M. (2019). *Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data*. 97, 302–311. <https://proceedings.mlr.press/v97/antelmi19a.html>
- Chadebec, C., Thibeau-Sutre, E., Burgos, N., & Allasonnière, S. (2023). Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 2879–2896. <https://doi.org/10.1109/TPAMI.2022.3185773>
- Dorent, R., Haouchine, N., Kogl, F., Joutard, S., Juvekar, P., Torio, E., Golby, A. J., Ourselin, S., Frisken, S., Vercauteren, T., Kapur, T., & Wells, W. M. (2023). Unified brain MR-ultrasound synthesis using multi-modal hierarchical representations. In *Medical image computing and computer assisted intervention – MICCAI 2023* (pp. 448–458). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-43999-5\\_43](https://doi.org/10.1007/978-3-031-43999-5_43)
- Hwang, H., Kim, G.-H., Hong, S., & Kim, K.-E. (2021). Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34, 12194–12207.
- Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. arXiv. <http://arxiv.org/abs/1312.6114>
- Lawry Aguila, A., Chapman, J., & Altmann, A. (2023). *Multi-modal variational autoencoders for normative modelling across multiple imaging modalities*. 425–434. [https://doi.org/10.1007/978-3-031-43907-0\\_41](https://doi.org/10.1007/978-3-031-43907-0_41)
- Lee, M., & Pavlovic, V. (2021). *Private-shared disentangled multimodal VAE for learning of latent representations*. 1692–1700. <https://doi.org/10.1109/CVPRW53098.2021.00185>
- Masahiro Suzuki, T. K., & Matsuo, Y. (2023). Pixyz: A python library for developing deep generative models. In *Advanced Robotics* (No. 0; Vol. 0, pp. 1–16). Taylor & Francis. <https://doi.org/10.1080/01691864.2023.2244568>
- Palumbo, E., Daunhawer, I., & Vogt, J. E. (2023). MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=sdQGxouELX>
- Palumbo, E., Manduchi, L., Laguna, S., Chopard, D., & Vogt, J. E. (2024). *Deep generative clustering with multimodal diffusion variational autoencoders*. <https://openreview.net/forum?id=k5THrhXDV3>
- Sejnova, G., Vavrecka, M., Stepanova, K., & Taniguchi, T. (2024). *Benchmarking multimodal variational autoencoders: CdSprites+ dataset and toolkit*. <https://arxiv.org/abs/2209.03048>
- Senellart, A., Chadebec, C., & Allasonnière, S. (2023). Improving multimodal joint variational autoencoders through normalizing flows and correlation analysis. *arXiv Preprint arXiv:2305.11832*.
- Shi, Y., Siddharth, N., Paige, B., & Torr, P. H. S. (2019). Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. *arXiv:1911.03393 [Cs, Stat]*. <http://arxiv.org/abs/1911.03393>
- Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2020). Multimodal generative learning utilizing jensen-shannon-divergence. *CoRR*, abs/2006.08242. <https://arxiv.org/abs/2006.08242>

- 122 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized Multimodal ELBO. *ICLR*.
- 123 Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Advanced*  
124 *Robotics*, 36(5-6), 261–278. <https://doi.org/10.1080/01691864.2022.2035253>
- 125 Suzuki, M., & Matsuo, Y. (2023). *Mitigating the limitations of multimodal VAEs with*  
126 *coordination-based approach*. <https://openreview.net/forum?id=Rn8u4MYgeNJ>
- 127 Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint Multimodal Learning with Deep  
128 Generative Models. *arXiv:1611.01891 [Cs, Stat]*. <http://arxiv.org/abs/1611.01891>
- 129 Tian, Y., & Engel, J. (2019). Latent Translation: Crossing Modalities by Bridging Generative  
130 Models. *ArXiv*.
- 131 Vasco, M., Yin, H., Melo, F. S., & Paiva, A. (2022). Leveraging hierarchy in multimodal  
132 generative models for effective cross-modality inference. *Neural Networks*, 146, 238–255.
- 133 Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2018). Generative Models of Visually  
134 Grounded Imagination. *arXiv:1705.10762 [Cs, Stat]*. <http://arxiv.org/abs/1705.10762>
- 135 Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation  
136 learning. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on*  
137 *machine learning* (Vol. 37, pp. 1083–1092). PMLR. [https://proceedings.mlr.press/v37/](https://proceedings.mlr.press/v37/wangb15.html)  
138 [wangb15.html](https://proceedings.mlr.press/v37/wangb15.html)
- 139 Wang, W., Yan, X., Lee, H., & Livescu, K. (2017). *Deep Variational Canonical Correlation*  
140 *Analysis*. <https://doi.org/10.48550/arXiv.1610.03454>
- 141 Wu, M., & Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised  
142 Learning. *Advances in Neural Information Processing Systems*, 31. [https://proceedings.](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)  
143 [neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)