

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

Agathe Senellart^{1,2,3}✉, Clément Chadebec^{1,2,3}, and Stéphanie Allasonnière^{1,2,3}

¹ Université de Paris-Cité ² Inria ³ 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Published: unpublished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, there has been a major boom in the development of multimodal machine learning models. Among open topics, representation (fusion) and generation of multimodal data are very active fields of research. Recently, Multimodal Variational Autoencoders (VAEs) have been attracting growing interest for both tasks, thanks to their versatility, scalability, and interpretability as probabilistic latent variable models. They are also particularly interesting in *partially observed* settings, as most can be trained even with missing data. This last point makes them particularly suited for the medical field, where available datasets are often incomplete ([Antelmi et al., 2019](#); [Lawry Aguila et al., 2023](#)).

We present MultiVae, an open-source Python library bringing together unified implementations of multimodal VAEs. It has been designed for easy and customizable use of these models on fully or partially observed data. This library also facilitates the development and benchmarking of new algorithms by integrating several benchmark datasets, a collection of evaluation metrics and tools for monitoring and sharing models.

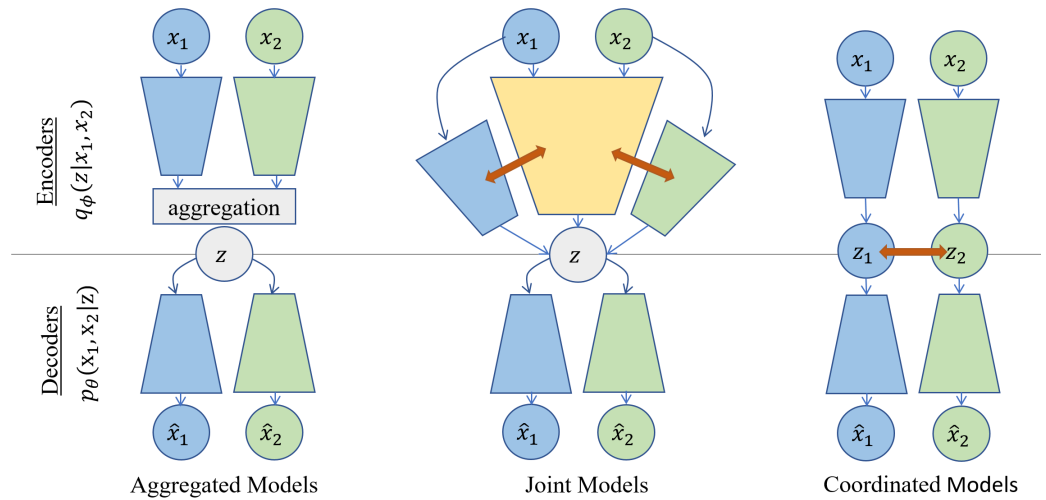
Multimodal Variational Autoencoders

Two main goals are commonly pursued in Multimodal Machine Learning: (1) Learn a shared representation from multiple modalities; (2) Learn to generate one missing modality given the ones that are available.

Multimodal Variational Autoencoders aim at solving both issues at the same time. These models learn a latent representation z of all modalities in a lower dimensional common 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 the 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 $q_\phi(z|X)$. As illustrated in the figure below, they can be categorized into three main groups: *Aggregated models* ([Shi et al., 2019](#); [Sutter et al., 2021](#); [Wu & Goodman, 2018](#)) use a mean or a product operation to aggregate the information coming from all modalities, where *Joint models* ([Senellart et al., 2023](#); [Suzuki et al., 2016](#); [Vedantam et al., 2018](#)) use a neural network taking all modalities as input. Finally *coordinated models* ([Tian & Engel, 2019](#); [Wang et al., 2017](#)) use different latent spaces while adding a constraint term in the loss to force them to be similar.



↔ : Similarity constraint

We designed our library MultiVae with the aim to implement all the approaches in a unified yet modular 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. However, except in our library MultiVae, there does not exist an implementation of these models that can be used on incomplete datasets in a straightforward manner. We propose a convenient way to handle missing modalities using *masks* in the loss computation of each aggregated model.

Data Augmentation

Another application of VAEs is Data Augmentation (DA): from sampling new latent codes z and decoding them with trained models, *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, the use of similar sampling techniques with multimodal VAEs remains largely unexplored. In our library, we provide a module `multivae.samplers` with popular sampling strategies to further explore the generative abilities of these models.

Statement of Need

Although multimodal VAEs have interesting applications in different fields, the lack of easy-to-use and verified implementations might hinder applicative research. With MultiVae, we offer unified implementations, designed to be accessible even for non-specialists. In order to provide reliable implementations, we reproduced, whenever possible, a key result from the original paper. Related software packages have grouped together model implementations: the [Multimodal VAE Comparison Toolkit](#) (Sejnova et al., 2024) includes 4 models and the [Pixyz](#) (Masahiro Suzuki & Matsuo, 2023) library contains 2 multimodal models. The most closely related work, released while we were developing our library, is `multi-view-ae` (Aguila et al., 2023), which contains a dozen of models. We compare in a summarizing table below, the different features of each work. Our library differs and complements existing software packages as follows: our API is quite different compared to previous work, the models implemented are not all the same, and for those we have in common, our implementation offers additional options. Indeed, for each model, we made sure to offer great flexibility on parameters' choices and to include all implementation details present in the original codes. Our library also offers

73 additional features: **compatibility with incomplete data**, which we consider essential for real-life
 74 applications, **samplers** to boost the generative abilities of models, and a range of tools dedicated
 75 to research and development such **benchmark datasets** and **metrics**. We implement the most
 76 commonly used metrics in a modular way to easily evaluate any model.

77 List of Models and Features

78 In the table below, we list available models and features, and compare to previous work.
 79 Symbol (✓*) indicates that the implementation includes additional options.

Models/ Features	Ours	(Aguila et al., 2023)	(Sejnova et al., 2024)
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	✓		

80 Code Quality and Documentation

81 Our code is available on Github (<https://github.com/AgatheSenellart/MultiVae>) and Pypi
 82 and we provide a full online documentation at (<https://multivae.readthedocs.io/>). Our code is
 83 unit-tested with a code coverage of 94%. The main features are illustrated through **tutorials**
 84 made available either as notebooks or scripts allowing users to get started easily. To further
 85 showcase how to use our library for research applications, we provide detailed *case studies* in
 86 the documentation.

Acknowledgements

We are grateful to the authors of all the initial implementations of the models included in MultiVae. This work benefited from state grant managed by the Agence Nationale de la Recherche under the France 2030 program, AN-23-IACL-0008. This research has been partly supported by the 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., Thibaud-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
- 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
- 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*

- 132 *arXiv:2305.11832*.
- 133 Shi, Y., Siddharth, N., Paige, B., & Torr, P. H. S. (2019). Variational Mixture-of-Experts
134 Autoencoders for Multi-Modal Deep Generative Models. *arXiv:1911.03393 [Cs, Stat]*.
135 <http://arxiv.org/abs/1911.03393>
- 136 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2020). Multimodal generative learning utilizing
137 jensen-shannon-divergence. *CoRR, abs/2006.08242*. <https://arxiv.org/abs/2006.08242>
- 138 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized Multimodal ELBO. *ICLR*.
- 139 Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Advanced*
140 *Robotics*, 36(5-6), 261–278. <https://doi.org/10.1080/01691864.2022.2035253>
- 141 Suzuki, M., & Matsuo, Y. (2023). *Mitigating the limitations of multimodal VAEs with*
142 *coordination-based approach*. <https://openreview.net/forum?id=Rn8u4MYgeNJ>
- 143 Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint Multimodal Learning with Deep
144 Generative Models. *arXiv:1611.01891 [Cs, Stat]*. <http://arxiv.org/abs/1611.01891>
- 145 Tian, Y., & Engel, J. (2019). Latent Translation: Crossing Modalities by Bridging Generative
146 Models. *ArXiv*.
- 147 Vasco, M., Yin, H., Melo, F. S., & Paiva, A. (2022). Leveraging hierarchy in multimodal
148 generative models for effective cross-modality inference. *Neural Networks*, 146, 238–255.
- 149 Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2018). Generative Models of Visually
150 Grounded Imagination. *arXiv:1705.10762 [Cs, Stat]*. <http://arxiv.org/abs/1705.10762>
- 151 Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation
152 learning. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on*
153 *machine learning* (Vol. 37, pp. 1083–1092). PMLR. [https://proceedings.mlr.press/v37/](https://proceedings.mlr.press/v37/wangb15.html)
154 [wangb15.html](https://proceedings.mlr.press/v37/wangb15.html)
- 155 Wang, W., Yan, X., Lee, H., & Livescu, K. (2017). *Deep Variational Canonical Correlation*
156 *Analysis*. <https://doi.org/10.48550/arXiv.1610.03454>
- 157 Wu, M., & Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised
158 Learning. *Advances in Neural Information Processing Systems*, 31. [https://proceedings.](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)
159 [neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)