

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

Agathe Senellart^{1,2*}¶ and Stéphanie Allasonnière^{2*}

1 Université de Paris-Cité 2 Inria 3 Inserm ¶ Corresponding author * These authors contributed equally.

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Published: unpublished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](#)).

In partnership with



This article and software are linked with research article DOI [10.3847/xxxxx](https://doi.org/10.3847/xxxxx) <- [update this with the DOI from AAS once you know it.](#), published in the *Astrophysical Journal* <- The name of the AAS journal..

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 models in the *partially observed* setting, as most of them can learn even with missing data. This last point makes them particularly interesting for research fields such as the medical field, where missing data are commonplace Lawry Aguila et al. (2023).

In this article, we present MultiVae, an open-source Python library for bringing together unified implementations of multimodal VAEs. It has been designed for easy, 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 variety of evaluation metrics and tools for monitoring and sharing models.

Multimodal Variational Autoencoders

In Multimodal Machine Learning, two goals are generally targeted: (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 any modality (Suzuki & Matsuo, 2022).

Let $X = (x_1, x_2, \dots, x_M)$ contain M modalities. In the VAE setting, we suppose that the generative process behind the observed data is the following:

$$z \sim p(z) \quad \forall 1 \leq i \leq M, x_i | z \sim p_\theta(x_i | z) \quad (1)$$

where $p(z)$ is a prior distribution that is often fixed, and $p_\theta(x_i | z)$ are called *decoders* and are parameterized by neural network. Typically, $p_\theta(x_i | z) = \mathcal{N}(x_i; \mu_\theta(z), \sigma_\theta(z))$ where $\mu_\theta, \sigma_\theta$ are neural networks. We aim to learn these *decoders* that translate z into the high dimensional data x_i . At the same time, we aim to learn an *encoder* $q_\phi(z | X)$ that map the multimodal observation to the latent space. $q_\phi(z | X)$ is also parameterized by a neural network. Derived from variational inference (Kingma & Welling, 2014), the VAE objective writes:

$$\mathcal{L}(X) = \mathbb{E}_{q_\phi(z|X)} \left(\sum_i \ln(p_\theta(x_i | z)) \right) - KL(q_\phi(z|X) | p(z))$$

The first term is a reconstruction loss and the second term can be seen as a regularization term that avoids overfitting. A typical training of a multimodal VAE consists in encoding the

data with the encoder, reconstructing each modality with the decoders and take a gradient step to optimize the loss $\mathcal{L}(X)$.

Most multimodal VAEs differ in how they construct the encoder $q_\phi(z|X)$. In the figure below, we summarize several approaches: *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., 2017a) uses different latent spaces but add a constraint term in the loss to force them to be similar.

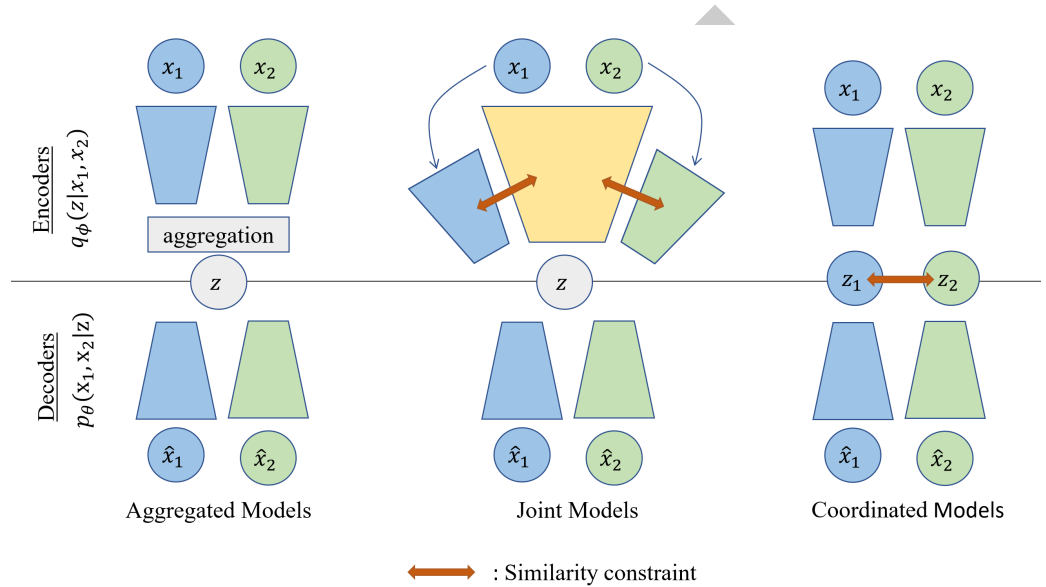


Figure 1: Different types of multimodal VAEs

Recent extensions of multimodal VAEs include additional terms to the loss, or use multiple (Palumbo et al., 2023) or hierarchical (Vasco et al., 2022; ?) latent spaces to more comprehensively describe the multimodal data. In our library, we implement all these approaches in an unified and modular way.

Aggregated models offer a natural way of learning on incomplete datasets: for an incomplete sample X , we use only the available modalities to encode the data and compute the loss $\mathcal{L}(X)$. However, except in MultiVae, there doesn't exist an implementation of these models that can be used on incomplete datasets in a straightforward manner.

Data Augmentation

Another application of these models is Data Augmentation (DA): from sampling latent codes z and decoding them, *fully synthetic multimodal* samples can be generated to augment a dataset. Data augmentation has been proven useful in many data-intensive deep learning applications (Chadebec et al., 2023). In a dedicated module `multivae.samplers`, we propose different ways of sampling latent codes z 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 easy to use by non-specialists and even on incomplete

data. To this end, we offer online documentation and tutorials. In order to propose reliable implementations of each method, we tried to reproduce, whenever possible, a key result from the original paper. Some works similar to ours 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 work closest to ours and 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 complements what already exists: 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 parameterization options. Indeed, for each model, we've made sure to offer great flexibility on parameters and to include all implementation details present in the original codes that boost results. What's more, our library offers many additional features: **compatibility with incomplete data**, which we consider essential for real-life applications, and a range of tools dedicated to the research and development of new algorithms: benchmark datasets, metrics modules and samplers, for testing and analyzing models. Our library also supports distributed training and straightforward model sharing via HuggingFace Hub (Face, 2023).

List of models and features

In the Table below, we list available models and features, and compare to previous work. This symbol (✓*) indicates that the implementation include 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)	✓	✓*	✓
MVTC AE (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., 2017b)		✓	
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	✓		
Ready-to-use Datasets	✓		✓
Model sharing via Hugging Face	✓		

An important difference in our user-interface, is that we handle all training and model parameters

84 within python dataclasses while (Sejnova et al., 2024; ?) uses independant YAML configuration
85 files.

86 Code quality and documentation

87 Our code is available on Github (<https://github.com/AgatheSenellart/MultiVae>) and Pypi and
88 we provide a full online documentation at (<https://multivae.readthedocs.io/en/latest/>). The
89 main features are illustrated through tutorials made available either as notebooks or scripts
90 allowing users to get started easily. To further showcase how to use our library for research
91 applications, we provide detailed case studies [here](#).

92 Acknowledgements

93 We are grateful to the authors of all the initial implementations of the models included in
94 MultiVae.

95 References

- 96 Aguila, A. L., Jayme, A., Montaña-Brown, N., Heuveline, V., & Altmann, A. (2023). Multi-
97 view-AE: A python package for multi-view autoencoder models. *Journal of Open Source*
98 *Software*, 8(85), 5093. <https://doi.org/10.21105/joss.05093>
- 99 Antelmi, L., Ayache, N., Robert, P., & Lorenzi, M. (2019). *Sparse multi-channel variational*
100 *autoencoder for the joint analysis of heterogeneous data*. 97, 302–311. [https://proceedings.](https://proceedings.mlr.press/v97/antelmi19a.html)
101 [mlr.press/v97/antelmi19a.html](https://proceedings.mlr.press/v97/antelmi19a.html)
- 102 Chadebec, C., Thibeau-Sutre, E., Burgos, N., & Allasonnière, S. (2023). Data augmentation
103 in high dimensional low sample size setting using a geometry-based variational autoencoder.
104 *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 2879–2896.
105 <https://doi.org/10.1109/TPAMI.2022.3185773>
- 106 Dorent, R., Haouchine, N., Kogl, F., Joutard, S., Juvekar, P., Torio, E., Golby, A. J., Ourselin,
107 S., Frisken, S., Vercauteren, T., Kapur, T., & Wells, W. M. (2023). Unified brain MR-
108 ultrasound synthesis using multi-modal hierarchical representations. In *Medical image*
109 *computing and computer assisted intervention – MICCAI 2023* (pp. 448–458). Springer
110 Nature Switzerland. https://doi.org/10.1007/978-3-031-43999-5_43
- 111 Face, H. (2023). *Hugging face hub*. <https://huggingface.co/docs/hub/index>
- 112 Hwang, H., Kim, G.-H., Hong, S., & Kim, K.-E. (2021). Multi-view representation learning
113 via total correlation objective. *Advances in Neural Information Processing Systems*, 34,
114 12194–12207.
- 115 Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. arXiv. [http:](http://arxiv.org/abs/1312.6114)
116 [//arxiv.org/abs/1312.6114](http://arxiv.org/abs/1312.6114)
- 117 Lawry Aguila, A., Chapman, J., & Altmann, A. (2023). *Multi-modal variational autoencoders*
118 *for normative modelling across multiple imaging modalities*. 425–434. [https://doi.](https://doi.org/10.1007/978-3-031-43907-0_41)
119 [org/10.1007/978-3-031-43907-0_41](https://doi.org/10.1007/978-3-031-43907-0_41)
- 120 Lee, M., & Pavlovic, V. (2021). *Private-shared disentangled multimodal VAE for learning of*
121 *latent representations*. 1692–1700. <https://doi.org/10.1109/CVPRW53098.2021.00185>
- 122 Masahiro Suzuki, T. K., & Matsuo, Y. (2023). Pixyz: A python library for developing deep
123 generative models. In *Advanced Robotics* (No. 0; Vol. 0, pp. 1–16). Taylor & Francis.
124 <https://doi.org/10.1080/01691864.2023.2244568>

- 125 Palumbo, E., Daunhawer, I., & Vogt, J. E. (2023). *MMVAE+: ENHANCING THE GENERA-*
126 *TIVE QUALITY OF MULTIMODAL VAES WITHOUT COMPROMISES.*
- 127 Palumbo, E., Manduchi, L., Laguna, S., Chopard, D., & Vogt, J. E. (2024). *Deep generative*
128 *clustering with multimodal diffusion variational autoencoders.* [https://openreview.net/](https://openreview.net/forum?id=k5THrhXDV3)
129 [forum?id=k5THrhXDV3](https://openreview.net/forum?id=k5THrhXDV3)
- 130 Sejnova, G., Vavrecka, M., Stepanova, K., & Taniguchi, T. (2024). *Benchmarking multimodal*
131 *variational autoencoders: CdSprites+ dataset and toolkit.* [https://arxiv.org/abs/2209.](https://arxiv.org/abs/2209.03048)
132 [03048](https://arxiv.org/abs/2209.03048)
- 133 Senellart, A., Chadebec, C., & Allasonnière, S. (2023). Improving multimodal joint varia-
134 tional autoencoders through normalizing flows and correlation analysis. *arXiv Preprint*
135 *arXiv:2305.11832.*
- 136 Shi, Y., Siddharth, N., Paige, B., & Torr, P. H. S. (2019). Variational Mixture-of-Experts
137 Autoencoders for Multi-Modal Deep Generative Models. *arXiv:1911.03393 [Cs, Stat].*
138 <http://arxiv.org/abs/1911.03393>
- 139 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2020). Multimodal generative learning utilizing
140 jensen-shannon-divergence. *CoRR, abs/2006.08242.* <https://arxiv.org/abs/2006.08242>
- 141 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized Multimodal ELBO. *ICLR.*
- 142 Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Advanced*
143 *Robotics*, 36(5-6), 261–278. <https://doi.org/10.1080/01691864.2022.2035253>
- 144 Suzuki, M., & Matsuo, Y. (2023). *Mitigating the limitations of multimodal VAEs with*
145 *coordination-based approach.* <https://openreview.net/forum?id=Rn8u4MYgeNJ>
- 146 Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint Multimodal Learning with Deep
147 Generative Models. *arXiv:1611.01891 [Cs, Stat].* <http://arxiv.org/abs/1611.01891>
- 148 Tian, Y., & Engel, J. (2019). Latent Translation: Crossing Modalities by Bridging Generative
149 Models. *ArXiv.*
- 150 Vasco, M., Yin, H., Melo, F. S., & Paiva, A. (2022). Leveraging hierarchy in multimodal
151 generative models for effective cross-modality inference. *Neural Networks*, 146, 238–255.
- 152 Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2018). Generative Models of Visually
153 Grounded Imagination. *arXiv:1705.10762 [Cs, Stat].* <http://arxiv.org/abs/1705.10762>
- 154 Wang, W., Yan, X., Lee, H., & Livescu, K. (2017b). *Deep Variational Canonical Correlation*
155 *Analysis.* arXiv. <https://doi.org/10.48550/arXiv.1610.03454>
- 156 Wang, W., Yan, X., Lee, H., & Livescu, K. (2017a). *Deep Variational Canonical Correlation*
157 *Analysis.* arXiv. <https://doi.org/10.48550/arXiv.1610.03454>
- 158 Wu, M., & Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised
159 Learning. *Advances in Neural Information Processing Systems*, 31. [https://proceedings.](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)
160 [neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html](https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html)