Masked Multi-time Diffusion for Multi-modal Generative Modeling

Mustapha Bounoua
Renault Software Factory
EURECOM, France
mustapha.bounoua@eurecom.fr

Giulio Franzese
Department of Data Science
EURECOM, France
giulio.franzese@eurecom.fr

Pietro Michiardi
Department of Data Science
EURECOM, France
pietro.michiardi@eurecom.fr

Abstract

Multi-modal data is ubiquitous, and models to learn a joint representation of all modalities have flourished. However, existing approaches suffer from a coherence-quality tradeoff, where generation quality comes at the expenses of generative coherence across modalities, and vice versa. To overcome these limitations, we propose a novel method that uses a set of independently trained, uni-modal, deterministic autoencoders. Individual latent variables are concatenated and fed to a masked diffusion model to enable generative modeling. We also introduce a new multi-time training method to learn the conditional score network for multi-modal diffusion. Empirically, our methodology substantially outperforms competitors in both generation quality and coherence.

1 Introduction

Multi-modal generative modelling is a crucial area of research in machine learning that aims to generate data according to multiple modalities, such as images, text, audio, and more. Indeed, real-world observations are often captured in various forms, and combining multiple modalities describing the same information can be an invaluable asset. For instance, images and text can provide complementary information in describing an object, audio and video can capture different aspects of a scene. Multi-modal generative models can also help in tasks such as data augmentation [9, 3, 28], missing modality imputation [2, 6, 46, 40], and conditional generation [14, 19].

Multi-modal models have flourished over the past years and have seen a tremendous interest from academia and industry, especially in the content creation sector. Whereas most recent approaches focus on specialization, by considering text as primary input to be associated mainly to images [26, 27, 25, 38, 44, 22, 5] and videos [4, 13, 31], in this work we target an established literature whose scope is more general, and in which all modalities are considered equally important. A large body of work rely on extensions of the VAE [18] to the multi-modal domain: initially interested in learning joint latent representation of multi-modal data, such works have mostly focused on generative modeling. Multi-modal generative models aim at *high-quality* data generation, as well as generative *coherence* across all modalities. These objectives apply to both joint generation of new data, and to conditional generation of missing modalities, given a disjoint set of available modalities.

In short, multi-modal Variational Autoencoders (VAEs) rely on combinations of uni-modal VAEs, and the design space consists mainly in the way the uni-modal latent variables are combined, to

Neural Information Processing Systems (NeurIPS) 2023 Workshop on Diffusion Models.

construct the joint posterior distribution. Early work such as [45] adopt a product of experts approach, whereas others [29] consider a mixture of expert approach. Product-based models achieve high generative quality, but suffer in terms of both joint and conditional coherence. This was found to be due to experts mis-calibration issues [29, 37]. On the other hand, mixture-based models produce coherent but qualitatively poor samples. A first attempt to address the so called **coherence-quality tradeoff** [7] is represented by the mixture of product of experts approach [37]. However recent comparative studies [7] show that none of the existing approaches fulfill both the generative quality and coherence criteria. A variety of techniques aim at finding a better operating point, such as contrastive learning techniques [30], hierarchical schemes [42], total correlation based calibration of single modality encoders [15], or different training objectives [36]. More recently, [24] considers explicitly separated shared and private latent spaces to overcome the aforementioned limitations. In this work, we propose a new method for multi-modal generative modeling that, by design, does not suffer from the aforementioned limitations, as supported by an extensive experimental campaign.

2 Our Method

Consider the random variable $X = \{X^1, \dots, X^M\} \sim p_D(x^1, \dots, x^M)$, consisting in the set of M of modalities sampled from the (unknown) multi-modal data distribution p_D. We indicate the marginal distribution of a single modality by $X^i \sim p^i_D(x^i)$ and the collection of a generic subset of modalities by $X^A \sim p^A_D(x^A)$, with $X^A \overset{\text{def}}{=} \{X^i\}_{i \in A}$, where $A \subset \{1, \dots, M\}$ is a set of indexes.

We use deterministic uni-modal autoencoders, whereby each modality X^i is encoded through its encoder $e^i_{\psi^i}$, into the modality specific latent variable Z^i and decoded into the corresponding $\hat{X}^i = d^i_{\theta^i}(Z^i)$. Since the mapping from input to latent is deterministic, there is no loss of information between X and Z.[1] Moreover, this choice avoids any form of interference in the back-propagated gradients corresponding to the uni-modal reconstruction losses. Consequently gradient conflicts issues [16], where stronger modalities pollute weaker ones, are avoided.

To enable such a simple design to become a generative model, we follow a two-stage approach [20, 39], where samples from the lower dimensional $q_\psi(z)$ are obtained through an appropriate generative model. We consider score-based diffusion models in latent space [26, 41] to solve this task, and call our approach Multi-modal Latent Diffusion (MLD) (see A for a schematic representation).

2.1 Multi-modal latent diffusion

In the first stage, the deterministic encoders project the input modalities X^i into the corresponding latent spaces Z^i. This transformation induces a distribution $q_\psi(z)$ for the latent variable $Z = [Z^1, \dots, Z^M]$, resulting from the concatenation of uni-modal latent variables.

Joint generation. To generate a new sample for all modalities we use a simple score-based diffusion model in latent space [32, 35, 41, 20, 39]. This requires reversing a stochastic noising process, starting from a simple, Gaussian distribution. Formally, the noising process is defined by a Stochastic Differential Equation (SDE) of the form:

$$\mathrm{d}R_t = \alpha(t) R_t \mathrm{d}t + g(t) \mathrm{d}W_t, \quad R_0 \sim q(r, 0), \tag{1}$$

where $\alpha(t)R_t$ and $g(t)$ are the drift and diffusion terms, respectively, and W_t is a Wiener process. The time-varying probability density $q(r, t)$ of the stochastic process at time $t \in [0, T]$, where T is finite, satisfies the Fokker-Planck equation [23], with initial conditions $q(r, 0)$. We assume uniqueness and existence of a stationary distribution $\rho(r)$ for the process Eq. (1).[2] The forward diffusion dynamics depend on the initial conditions $R_0 \sim q(r, 0)$. We consider $R_0 = Z$ to be the initial condition for the diffusion process, which is equivalent to $q(r, 0) = q_\psi(r)$. Under loose conditions [1], a time-reversed stochastic process exists, with a new SDE of the form:

$$\mathrm{d}R_t = \left(-\alpha(T-t)R_t + g^2(T-t)\nabla \log(q(R_t, T-t))\right)\mathrm{d}t + g(T-t)\mathrm{d}W_t, \quad R_0 \sim q(r, T), \tag{2}$$

indicating that, in principle, simulation of Eq. (2) allows to generate samples from the desired distribution $q(r, 0)$. In practice, we use a **parametric score network** $s_\chi(r, t)$ to approximate the true

[1] Since the measures are not absolutely continuous w.r.t the Lebesgue measure, mutual information is $+\infty$.

[2] This is not necessary for the validity of the method [34]

score function, and we approximate $q(r,T)$ with the stationary distribution $\rho(r)$. The joint generation of all modalities is achieved through the simulation of the reverse-time SDE in Eq. (2) to obtain $Z \sim q_\psi(z)$ which can be decoded into samples using the deterministic decoders.

Conditional generation. Given a generic partition of all modalities into non overlapping sets $A_1 \cup A_2$, where $A_2 = (\{1,\dots,M\} \setminus A_1)$, conditional generation requires samples from the conditional distribution $q_\psi(z^{A_1} \mid z^{A_2})$, which are based on *masked* forward and backward diffusion processes. Given conditioning latent modalities z^{A_2}, we consider a modified forward diffusion process with initial conditions $R_0 = \mathcal{C}(R_0^{A_1}, R_0^{A_2})$, with $R_0^{A_1} \sim q_\psi(r^{A_1} \mid z^{A_2})$, $R_0^{A_2} = z^{A_2}$. The composition operation $\mathcal{C}(\cdot)$ concatenates generated (R^{A_1}) and conditioning latents (z^{A_2}). More formally, we define the masked forward diffusion SDE:

$$\mathrm{d}R_t = m(A_1) \odot \left[\alpha(t) R_t \mathrm{d}t + g(t) \mathrm{d}W_t\right], \quad q(r,0) = q_\psi(r^{A_1} \mid z^{A_2}) \delta(r^{A_2} - z^{A_2}). \tag{3}$$

The mask $m(A_1)$ contains M vectors u^i, one per modality, and with the corresponding cardinality. If modality $j \in A_1$, then $u^j = \mathbf{1}$, otherwise $u^j = \mathbf{0}$. Then, the effect of masking is to "freeze" throughout the diffusion process the part of the random variable R_t corresponding to the conditioning latent modalities z^{A_2}. To sample from $q_\psi(z^{A_1} \mid z^{A_2})$, we derive the reverse-time dynamics of Eq. (3) as follows:

$$\mathrm{d}R_t = m(A_1) \odot \left[\left(-\alpha(T-t) R_t + g^2(T-t) \nabla \log\big(q(R_t, T-t \mid z^{A_2})\big)\right) \mathrm{d}t + g(T-t) \mathrm{d}W_t\right], \tag{4}$$

with initial conditions $R_0 = \mathcal{C}(R_0^{A_1}, z^{A_2})$ and $R_0^{A_1} \sim q(r^{A_1}, T \mid z^{A_2})$. Then, we approximate $q(r^{A_1}, T \mid z^{A_2})$ by its corresponding steady state distribution $\rho(r^{A_1})$, and the true (conditional) score function $\nabla \log\big(q(r,t \mid z^{A_2})\big)$ by a conditional score network $s_\chi(r^{A_1}, t \mid z^{A_2})$. A naïve approach would be to rely on the unconditional score network $s_\chi(r,t)$, by casting it as an *in-painting* objective. In the context of multi-modal data, the assumptions underlying in-painting are difficult to satisfy, because every modality exhibits different dynamics when perturbed by noise. Then, we propose a mechanism to learn the conditional score for any subset of conditioning and generated modalities.

2.2 Multi-time Diffusion

Instead of training a separate score network for each possible combination of conditional modalities, which is computationally infeasible, we use a single architecture that accepts all modalities as inputs and a *multi-time vector* $\tau = [t_1, \dots, t_M]$. The multi-time vector serves two purposes: it is both a conditioning signal and the time at which we observe the diffusion process.

Training: learning the conditional score network relies on randomization. As discussed in § 2.1, we consider an arbitrary partitioning of all modalities in two disjoint sets, A_1 and A_2. The set A_2 contains randomly selected conditioning modalities, while the remaining modalities belong to set A_1. Then, during training, the parametric score network estimates $\nabla \log\big(q(r,t \mid z^{A_2})\big)$, whereby the set A_2 is randomly chosen at every step. This is achieved by the *masked diffusion process* from Eq. (3), which only diffuses modalities in A_1. More formally, the score network input is $R_t = \mathcal{C}(R_t^{A_1}, Z^{A_2})$, along with a multi-time vector $\tau(A_1, t) = t\,[\mathbb{1}(1 \in A_1), \dots, \mathbb{1}(M \in A_1)]$.

Conditional generation: any valid numerical integration scheme for Eq. (4) can be used for conditional sampling (see A for Euler-Maruyama integrator pseudo-code). First, conditioning modalities in the set A_2 are encoded into the corresponding latent variables $z^{A_2} = \{e^j(x^j)\}_{j \in A_2}$. Then, numerical integration is performed with step-size $\Delta t = {}^{T}\!/_{N}$, starting from the initial conditions $R_0 = \mathcal{C}(R_0^{A_1}, z^{A_2})$, with $R_0^{A_1} \sim \rho(r^{A_1})$. At each integration step, the score network s_χ is fed the current state of the process and the multi-time vector $\tau(A_1, \cdot)$. Before updating the state, the masking is applied. Finally, the generated modalities are obtained thanks to the decoders.

3 Experiments

We compare our method MLD to MVAE [45], MMVAE [29], MOPOE [37], NEXUS [42] and MVTCAE [15], re-implementing competitors in the same code base as our method, and selecting their best hyper-parameters (as indicated by the authors). For fair comparison, we use the same encoder/decoder architecture for all the models (see C for more details). For MLD, the score network is implemented using a simple stacked MLP with skip connections (see A for more details).

Results. Overall, MLD largely outperforms alternatives from the literature, **both** in terms of coherence and generative quality. VAE-based models suffer from a coherence-quality trad-off and modality collapse for highly heterogeneous data-sets.

The first data-set we consider is **MNIST-SVHN** ([29]), where the two modalities differ in complexity. High variability, noise and ambiguity makes attaining good coherence for the SVHN modality a challenging task. Overall, MLD outperforms all VAE-based alternatives in terms of coherency, especially in terms of joint generation and conditional generation of MNIST given SVHN, see Table 1. Mixture models (MMVAE, MOPOE) suffer from modality collapse (poor SVHN generation), whereas product of experts (MVAE, MVTCAE) generate better quality samples at the expense of SVHN to MNIST conditional coherence. Joint generation is poor for all VAE models. MLD achieves the best performance also in terms of generation quality, as confirmed also by qualitative results (Figure 1) showing for example how MLD conditionally generates multiple SVHN digits within one sample, given the input MNIST image, whereas other methods fail to do so.

Table 1: Generation coherence and quality for **MNIST-SVHN** (M :MNIST, S: SVHN). *Quality* is measured in terms of FMD for MNIST and FID for SVHN. *Coherence* is measured as in [29, 37, 24], using pre-trained classifiers. Full details on the metrics are included in Appendix B. All results are averaged over 5 seeds.

Models	Coherence (%↑)			Quality (↓)			
	Joint	M → S	S → M	Joint(M)	Joint(S)	M → S	S → M
MVAE	38.19	48.21	28.57	13.34	68.9	68.0	13.66
MMVAE	37.82	11.72	67.55	25.89	146.82	393.33	53.37
MOPOE	39.93	12.27	68.82	20.11	129.2	373.73	43.34
NEXUS	40.0	16.68	70.67	13.84	98.13	281.28	53.41
MVTCAE	48.78	81.97	49.78	12.98	**52.92**	69.48	13.55
MLD	**85.22**	**83.79**	**79.13**	**3.93**	56.36	**57.2**	**3.67**

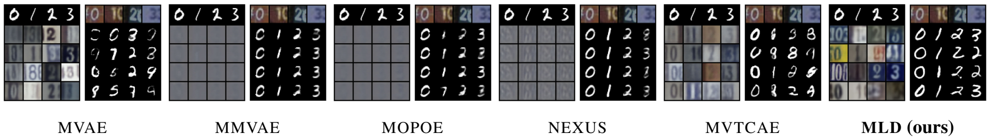

MVAE MMVAE MOPOE NEXUS MVTCAE **MLD (ours)**

Figure 1: Qualitative results for **MNIST-SVHN**. For each model we report: MNIST to SVHN conditional generation in the left, SVHN to MNIST conditional generation in the right.

Finally, we explore the Caltech Birds **CUB** [29] data-set, following the same experimentation protocol in [7] by using real bird images (instead of ResNet-features as in [29]). Figure 2 presents qualitative results for caption to image conditional generation. MLD is the only model capable of generating bird images with convincing coherence. Clearly, none of the VAE-based methods is able to achieve sufficient caption to image conditional generation quality using the same simple autoencoder architecture. Note that an image autoencoder with larger capacity improves considerably MLD generative performance, suggesting that careful engineering applied to modality specific autoencoders is a promising avenue for future work.

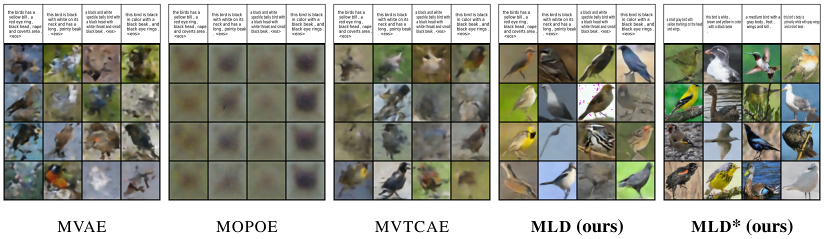

MVAE MOPOE MVTCAE **MLD (ours)** **MLD* (ours)**

Figure 2: Qualitative results on **CUB** data-set. Caption used as condition to generate images. **MLD*** denotes the version of our method using a more powerful image autoencoder for image modality.

4 Conclusion and Limitations

We have addressed the challenge of multi-modal generative modeling by proposing a novel method, Multi-modal Latent Diffusion (MLD). Our approach overcomes the coherence-quality tradeoff that is inherent in existing multi-modal VAE-based model. MLD, uses a set of independently trained, uni-modal, deterministic autoencoders. Generative properties of our model stem from a multi-time masked diffusion process that operates on latent variables and allows joint and conditonal generation. An extensive experimental campaign provided compelling evidence on the effectiveness of MLD for multi-modal generative modeling.

Acknowledgment

GF gratefully acknowledges support from the European Commission (ADROIT6G Grant agreement ID: 101095363).

References

[1] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[2] L. Antelmi, N. Ayache, P. Robert, and M. Lorenzi. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 302–311. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/antelmi19a.html.

[3] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023.

[4] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023.

[5] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan. Muse: Text-to-image generation via masked generative transformers, 2023.

[6] M. Da Silva–Filarder, A. Ancora, M. Filippone, and P. Michiardi. Multimodal variational autoencoders for sensor fusion and cross generation. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1069–1076, 2021. doi: 10.1109/ICMLA52953.2021.00175.

[7] I. Daunhawer, T. M. Sutter, K. Chin-Cheong, E. Palumbo, and J. E. Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w-CPUXXrAj.

[8] E. Dupont, H. Kim, S. M. A. Eslami, D. J. Rezende, and D. Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5694–5725. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/dupont22a.html.

[9] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. QI. IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nUmCcZ5RKF.

[10] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[12] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

[14] X. Huang, A. Mallya, T.-C. Wang, and M.-Y. Liu. Multimodal conditional image synthesis with product-of-experts gans. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, page 91–109, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19786-4. doi: 10.1007/978-3-031-19787-1_6. URL https://doi.org/10.1007/978-3-031-19787-1_6.

[15] H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34:12194–12207, 2021.

[16] A. Javaloy, M. Meghdadi, and I. Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/javaloy22a.html.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

[19] S. Lee, J. Ha, and G. Kim. Harmonizing maximum likelihood with GANs for multimodal conditional generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJxyAjRcFX.

[20] G. Loaiza-Ganem, B. L. Ross, J. C. Cresswell, and A. L. Caterini. Diagnosing and fixing manifold overfitting in deep generative models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=0nEZCVshxS.

[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[22] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[23] B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[24] E. Palumbo, I. Daunhawer, and J. E. Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sdQGxouELX.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

[28] M. B. Sariyildiz, K. Alahari, D. Larlus, and Y. Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones, 2023.

[29] Y. Shi, S. N, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0ae775a8cb3b499ad1fca944e6f5c836-Paper.pdf.

[30] Y. Shi, B. Paige, P. Torr, and S. N. Relating by contrasting: A data-efficient framework for multimodal generative models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=vhKe9UFbrJo.

[31] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[32] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[33] Y. Song and S. Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

[34] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.

[35] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[36] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *CoRR*, abs/2006.08242, 2020. URL https://arxiv.org/abs/2006.08242.

[37] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=5Y21V0RDBV.

[38] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu. Df-gan: A simple and effective baseline for text-to-image synthesis, 2022.

[39] B.-H. Tran, S. Rossi, D. Milios, P. Michiardi, E. V. Bonilla, and M. Filippone. Model selection for bayesian autoencoders. *Advances in Neural Information Processing Systems*, 34:19730–19742, 2021.

[40] L. Tran, X. Liu, J. Zhou, and R. Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[41] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=P9TYG0j-wtG.

[42] M. Vasco, H. Yin, F. S. Melo, and A. Paiva. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2 2022. ISSN 18792782. doi: 10.1016/j.neunet.2021.11.019.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[44] F. Wu, L. Liu, F. Hao, F. He, and J. Cheng. Text-to-image synthesis based on object-guided joint-decoding transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18122, June 2022.

[45] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1102a326d5f7c9e04fc3c89d0ede88c9-Paper.pdf.

[46] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou. Unified multi-modal image synthesis for missing modality imputation, 2023.

A Appendix

Multi-modal Latent Diffusion — Supplementary material

A Diffusion in the multimodal latent space

In this section, we provide additional technical details of Multi-modal Latent Diffusion (MLD).

A.1 Modalities Auto-Encoders

Each deterministic autoencoders used in the first stage of MLD uses a vector latent space with no size constraints. Instead, VAE-based models, generally require the latent space of each individual VAE to be exactly of the same size, to allow the definition of a joint latent space.

In our approach, before concatenation, the modality-specific latent spaces are *normalized* by element-wise mean and standard deviation. In practice, we use the statistics retrieved from the first training batch, which we found sufficient to gain sufficient statistical confidence. This operation allows the harmonization of different modality-specific latent spaces and, therefore, facilitate the learning of a joint score network.

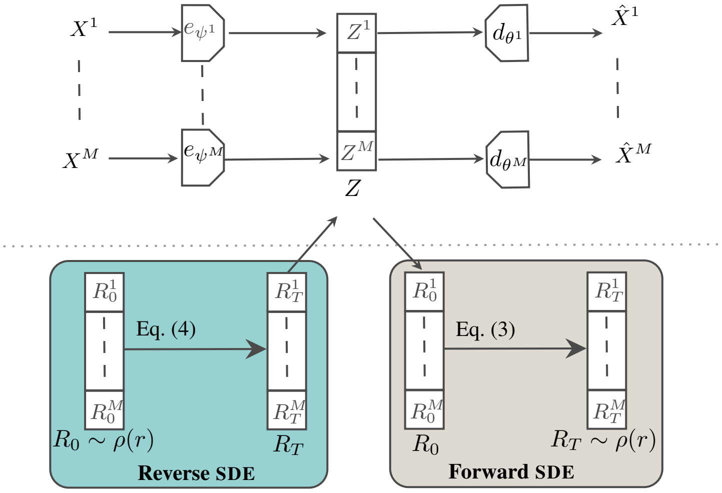


Figure 3: Multi-modal Latent Diffusion. Two-stage model involving: ***Top:*** deterministic, modality-specific encoder/decoders, ***Bottom:*** score-based diffusion model on the concatenated latent spaces.

A.2 Multi-modal latent diffusion

In § 2.1, we presented our multi-modal latent diffusion process allowing multi-modal joint and conditional generation. The role of the SDE is to gradually add noise to the data, perturbing its structure until attaining a noise distribution. In this work, we consider Variance preserving SDE (VPSDE) [35]. In this framework we have : $\rho(r) \sim \mathcal{N}(0; I)$, $\alpha(t) = -\frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$, where $\beta(t) = \beta_{min} + t(\beta_{max} - \beta_{min})$. Following [12, 35], we set $\beta_{min} = 0.1$ and $\beta_{max} = 20$. With this configuration and by substitution of Eq. (1), we obtain the following forward SDE:

$$\mathrm{d}R_t = -\frac{1}{2}\beta(t)R_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t, \qquad t \in [0, T]. \tag{5}$$

The corresponding perturbation kernel is given by :

$$q(r|z,t) = \mathcal{N}(r; e^{-\frac{1}{4}t^2(\beta_{max}-\beta_{min})-\frac{1}{2}t\beta_{min}}z, (1 - e^{-\frac{1}{2}t^2(\beta_{max}-\beta_{min})-t\beta_{min}})\mathbf{I}). \tag{6}$$

The marginal score $\nabla \log q(R_t, t)$ is approximated by a score network $s_\chi(R_t, t)$ whose parameters χ can be optimized by minimizing the following evidence lower bound (ELBO):

$$\mathrm{KL}[q_\psi(r)\,|\,|\,q(r,0)] \leq \frac{1}{2}\int\limits_0^T g^2(t)\mathbb{E}[\|s_\chi(R_t,t) - \nabla \log q(R_t,t)\|^2]\mathrm{d}t + KL[q(r,T)||\rho(r)], \quad (7)$$

where the first term on the r.h.s is referred to as score-matching objective, and is the loss over which the score network is optimized, and the second is a vanishing term for $T \to \infty$. We found that using the same re-scaling as in [35] is more stable.

The reverse process is described by a different SDE (Eq. (2)). When using a variance-preserving SDE, Eq. (2) specializes in:

$$\mathrm{d}R_t = \left[\frac{1}{2}\beta(T-t)R_t + \beta(T-t)\nabla \log q(R_t, T-t)\right]\mathrm{d}t + \sqrt{\beta(T-t)}\mathrm{d}W_t, \quad (8)$$

With $R_0 \sim \rho(r)$ as initial condition and time t flows from $t = 0$ to $t = T$.

Once the parametric score network is optimized, trough the simulation of Eq. (8), sampling $R_T \sim q_\psi(r)$ is possible allowing **joint generation**. A numerical SDE solver can be used to sample R_T which can be fed to the modality specific decoders to jointly sample a set of $\hat{X} = \{d^i_\theta(R^i_T)\}^M_{i=0}$.

A.3 Multi-time Diffusion

To learn the score network capable of both conditional and joint generation, we proposed in § 2 a multi-time masked diffusion process. Algorithm 1 presents a pseudo-code for the multi time masked training. The masked diffusion process is applied following a randomization with probably d.

At each step, a set of conditioning modalities A_2 is sampled from a predefined distribution ν, where $\nu(\emptyset) \overset{\text{def}}{=} \Pr(A_2 = \emptyset) = d$, and $\nu(U) \overset{\text{def}}{=} \Pr(A_2 = U) = {}^{(1-d)}/_{(2^M-1)}$ with $U \in \mathcal{P}(\{1, \dots, M\}) \setminus \emptyset$, where $\mathcal{P}(\{1, \dots, M\})$ is the powerset of all modalities. The remaining set of modalities A_1 is selected to be the diffused modalities. The time t is sampled uniformly from $[0, T]$ and the portion of the latent space corresponding to the subset A_1 is diffused accordingly. Using the masking as shown in Algorithm 1, the portion of the latent space corresponding to the subset A_2 is not diffused and forced to be equal to $R_0^{A_2} = z^{A_2}$. The multi-time vector τ is constructed. Lastly, the score network is optimized by minimizing a masked loss corresponding to the diffused part of the latent space. With probability $(1-d)$, $A_2 = \emptyset$ and all the modalities are diffused at the same time. In order to calibrate the loss, given that the randomization of A_1 and A_2 can result in diffusing different sizes of the latent space, we re-weight the loss according to the cardinality of the diffused and freezed portions of the latent space:

$$\Omega(A_1, A_2) = 1 + \frac{dim(A_2)}{dim(A_1)} \quad (9)$$

Where $\dim(.)$ is the sum of each latent space cardinality of a given subset of modalities with $dim(\emptyset) = 0$.

The optimized score network can approximate both the conditional and unconditional true score:

$$s_\chi(R_t, \tau(A_1, t)) \sim \nabla \log q(R_t, t \,|\, z^{A_2}). \quad (10)$$

The joint generation is a special case of the latter with $A_2 = \emptyset$:

$$s_\chi(R_t, \tau(A_1, t)) \sim \nabla \log q(R_t, t) \quad , A_1 = \{1, ..., M\} \quad (11)$$

Algorithm 2 describes the reverse conditional generation pseudo-code. One can generate a set of modalities A_1 conditioned on the available set of modalities A_2. First, the available modalities are encoded into their respective latent space z^{A_2} , the initial missing part is sampled from the stationary distribution $R_0^{A_1} \sim \rho(r^{A_1})$, using an SDE solver (e.g. Euler-Maruyama), the reverse diffusion SDE

Algorithm 1: MLD Masked Multi-time diffusion training step

```
**Data:** $X = \{x^i\}_{i=1}^M$
**Param:** $d$
$Z \leftarrow \{e_{\phi_i}(x^i)\}_{i=0}^M$          // Encode the modalities $X$ into their latent space
$A_2 \sim \nu$                                       // $\nu$ depends on the parameter $d$
$A_1 \leftarrow \{1, \dots, M\} \setminus A_2$
$t \sim \mathcal{U}[0, T]$
$R \sim q(r|Z, t)$      // Diffuse the available portion of the latent space(Eq. (6))
$R \leftarrow m(A_1) \odot R + (1 - m(A_1)) \odot Z$                      // Masked diffusion
$\tau(A_1, t) \leftarrow [\mathbb{1}(1 \in A_1)t, \dots, \mathbb{1}(M \in A_1)t]$       // Construct the multi time vector
**Return** $\nabla_\chi \left\{ \Omega(A_1, A_2) \quad \| m(A_1) \odot \quad \left[ s_\chi(R, \tau(A_1, t)) - \nabla \log q(R, t|z^{A_2}) \right] \|_2^2 \quad \right\}$
```

(in Eq. (8)) is discretized using a finite time steps $\Delta t = T/N$, starting from $t = 0$ and iterating until $t \approx T$. At each iteration , the score network s_χ is fed the current state of the process and the multi-time vector $\tau(A_1, \cdot)$. Lastly, the reverse diffusion update is applied only on the missing modality latent space . This process is repeated until obtaining $R_T^{A_1} = \hat{Z}^{A_1}$ which can be decoded to recover $\hat{x}^{A_1}$.

Algorithm 2: MLD conditional generation.

```
**Data:** $x^{A_2} \leftarrow \{x^i\}_{i \in A_2}$
$z^{A_2} \leftarrow \{e_{\phi_i}(x^i)\}_{i \in A_2}$      // Encode the available modalities $X$ into their latent
  space
$A_1 \leftarrow \{1, \dots, M\} \setminus A_2$                       // The set of modalities to be generated
$R_0 \leftarrow \mathcal{C}(R_0^{A_1}, z^{A_2}), \qquad R_0^{A_1} \sim \rho(r^{A_1})$          // Compose the initial latent space
$R \leftarrow R_0$
$\Delta t \leftarrow T/N$
**for** $n = 0$ ***to*** $N - 1$ **do**
    $t' \leftarrow T - n\,\Delta t$
    $\tau(A_1, t') \leftarrow [\mathbb{1}(1 \in A_1)t', \dots, \mathbb{1}(M \in A_1)t']$  // Construct the multi-time vector
    $\epsilon \sim \mathcal{N}(0; I)$ **if** $n < N$ **else** $\epsilon = 0$
    $\Delta R \leftarrow \Delta t \left[ \frac{1}{2}\beta(t')R + \beta(t')s_\chi(R, \tau(A_1, t')) \right] + \sqrt{\beta(t')\Delta t}\epsilon$
    $R \leftarrow R + \Delta R$                               // The Euler-Maruyama update step
    $R \leftarrow m(A_1) \odot R + (1 - m(A_1)) \odot R_0$   // Update the portion corresponding to
      the unavailable modalities
**end**
$\hat{z}^{A_1} = R^{A_1}$
**Return** $\hat{X}^{A_1} = \{d_\theta^i(\hat{z}^i)\}_{i \in A_1}$
```

A.4 Technical details

Sampling schedule: We use the sampling schedule proposed in [21], which has shown to improve the coherence of the conditional and joint generation. We use the best parameters suggested by the authors: $N = 250$ time-steps, applied $r = 10$ re-sampling times with jump size $j = 10$. For readability in Algorithm 2, we present pseudo code with a linear sampling schedule which can be easily adapted to any other schedule.

Training the score network: Inspired by the architecture from [8], we use simple Residual multilayer perceptron (MLP) blocks with skip connections as our score network (see Figure 4). We fix the **width** and **number of blocks** proportionally to the number of the modalities and the latent space size. As in [33], we use Exponential moving average (EMA) of model parameters with a momentum parameter $m = 0.999$.

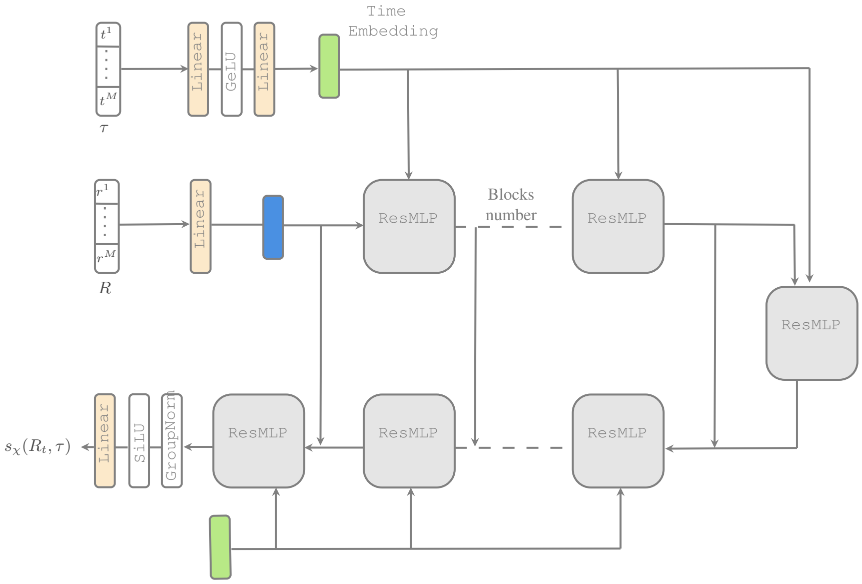


Figure 4: Score network s_χ architecture used in our MLD implementation. Residual MLP block architecture is shown in Figure 5.

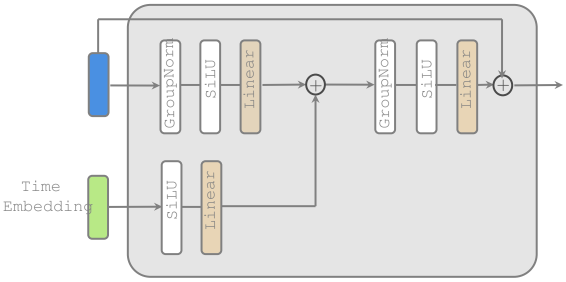


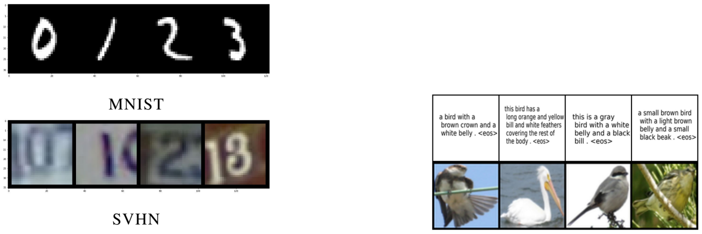
Figure 5: Architecture of ResMLP block.

B Datasets and evaluation protocol

B.1 Datasets description

MNIST-SVHN [29] is constructed using pairs of MNIST and SVHN, sharing the same digit class (See Figure 6a). Each instance of a digit class (in either dataset) is randomly paired with 20 instances of the same digit class from the other data-set. SVHN modality samples are obtained from house numbers in Google Street View images, characterized by a variety of colors, shapes and angles. A high number of SVHN samples are noisy and can contain different digits within the same sample due to the imperfect cropping of the original full house number image. One challenge of this data-set for multi-modal generative models is to learn to extract digit number and reconstruct a coherent MNIST modality.

CUB [29] is comprised of bird images and their associated text captions. The work in [29] used a simplified version based on pre-computed ResNet-features. We follow [7] and conduct all our experiments on the real image data instead. Each image from the 11,788 photos of birds from Caltech-Birds [43] are resized to $3 \times 64 \times 64$ image size and coupled with 10 textual descriptions of the respective bird (See Figure 6b).


MNIST
SVHN
a bird with a brown crown and a white belly . <eos>
this bird has a long orange and yellow bill and white feathers covering the rest of the body . <eos>
this is a gray bird with a white belly and a black bill . <eos>
a small brown bird with a light brown belly and a small black beak . <eos>

(a) MNIST-SVHN

(b) CUB

Figure 6: Illustrative example of the Datasets used for the evaluation

B.2 Evaluation metrics

Multimodal generative models are evaluated in terms of generative coherence and quality.

B.2.1 Generation Coherence

We measure *coherence* by verifying that generated data (both for joint and conditional generations) share the same information across modalities. Following [29, 37, 15, 42, 7], we consider the class label of the modalities as the shared information and use pre-trained classifiers to extract the label information form the generated samples and compare it across modalities.

For **MNIST-SVHN**, the shared semantic information is the digit class number. Single modality classifiers are trained to classify the digit number of a given modality sample. To compute the conditional generation of modality m with a subset of modalities A, we feed the modality specific pre-trained classifier $\mathbf{C}_m$ with the conditional generated sample $\hat{X}^m$. The predicted label class is compared to the ground truth label y_{X^A} which is the label of modalities of the subset X^A. For N samples, the matching rate average establishes the coherence. For all the experiments, N is equal to the length of the test-set.

$$
\textit{Coherence}(\hat{X}^m|X^A) = \frac{1}{N}\sum_{1}^{N} \mathbb{1}_{\{\mathbf{C}_m(\hat{X}^m)=y_{X^A}\}} \tag{12}
$$

The **joint generation coherence** is measured by feeding the generated samples of each modality to their specific trained classifier. The rate with which all classifiers output the same predicted digit label for N generations is considered as the joint generation coherence.

Due to the unavailability of labels in the **CUB** data-set, we use CLIP-Score (CLIP-S) [10] a state of the art metric for image captioning evaluation.

B.2.2 Generation Quality

For each modality, we consider the following metrics:

- **RGB Images**: Fréchet Inception Distance (FID) [11] is the state-of-the-art standard metric to evaluate image generation quality of generative models.
- **Other modalities** For other modality types, we derive Fréchet Modality Distance (FMD) (Fréchet Modality Distance), a similar metric to FID. We compute the **Fréchet distance** between the statistics retrieved from the activations of the modality specific pre-trained classifiers used for coherence evaluation. FMD is used to evaluate the generative quality of MNIST modality in **MNIST-SVHN**

For conditional generation, we compute the quality metric (FID,Fréchet Audio Distance (FAD) or FMD) using the conditionally generated modality and the real data. For joint generation, we use the randomly generated modality and randomly selected same number of samples from the real data.

For **CUB**, we use 10000 samples to evaluate the generation quality in terms of FID. In the remaining experiments, we use 5000 samples to evaluate the performance in terms of FID, FAD or FMD.

C Implementation details

We report in this section the implementation details for each benchmark. We used the same unified code-base for all the baselines, using the *PyTorch* framework. The VAE implementation is adapted from the official code whenever it's available (Product of Experts (MVAE), Mixture of Expert (MMVAE) and Mixture of Product of Experts (MOPOE) as in [3], Multi-view Total Correlation Autoencoder (MVTCAE) [4] and Hierarchical Genertive Model (NEXUS)[5]). For fairness, MLD and all the VAE-based models use the same autoencoder architecture. We use the best hyper-parameters suggested by the authors. Across all the data-sets, we use the *Adam optimizer* [17] for training.

C.1 MLD

MLD uses the same autoencoders architecture used for VAE-based models, except that these are deterministic autoencoders. The autoencoders are trained using the same reconstruction loss term as for the VAE-based models. Table 2 and Table 3 summarize the hyper-parameters used during the two phases of MLD training. Note that for the image modality in the CUB dataset, to overcome over-fitting in training the deterministic autoencoder, data augmentation was necessary (we used *TrivialAugmentWide* from the Torchvision library).

Table 2: MLD: The deterministic autoencoders hyper-parameters

Dataset	Modality	Latent space	Batch size	Lr	Epochs	Weight decay
MNIST-SVHN	MNIST	16	128	1e-3	150	
	SVHN	64				
CUB	Caption	32	128	1e-3	500	
	Image	64		1e-4	300	1e-6

Table 3: MLD: The score network hyper-parameters

Dataset	d	Blocks	Width	Time embed	Batch size	Lr	Epochs
MNIST-SVHN	0.5	2	512	256	128	1e-4	150
CUB	0.7	2	1024	512	64		3000

C.2 VAE-based models

For **MNIST-SVHN**, we follow [37, 29] and use the same autoencoder architecture and pre-trained classifier.The latent space size is set to 20, $\beta = 5.0$. For MVTCAE $\alpha = \frac{5}{6}$. For both modalities, the likelihood is estimated using Laplace distribution. For NEXUS, we use the same modalities latent space sizes as in MLD, the joint NEXUS latent space is set to 20, $\beta_i = 1.0$ and $\beta_c = 5.0$. We train all the VAE-models for 150 epochs with 256 batch size and learning rate of $1e - 3$.

For **CUB**, we use the same autoencoders architecture and implementation settings as in [7]. Laplace and one-hot categorical distributions are used to estimate likelihoods of the image and caption modalities respectively. The latent space size is set to 64, $\beta = 9.0$ for MVAE, MVTCAE and MOPOE and $\beta = 1$ for MMVAE. We set $\alpha = \frac{5}{6}$ for MVTCAE. For NEXUS, we use the same modalities latent space sizes as in MLD, the joint NEXUS latent space is set to 64, $\beta_i = 1.0$ and $\beta_c = 1$. We train all the models for 150 epochs with 64 batch size, with learning rate of $5e - 4$ for MVAE, MVTCAE and MOPOE and $1e - 3$ for the remaining models.

Finally, note that in the official implementation of [37] and [15], for the **MNIST-SVHN** data-sets, the classifiers were used for evaluation using dropout. In our implementation, we make sure to deactivate dropout during evaluation step.

[3] https://github.com/thomassutter/MoPoE

[4] https://github.com/gr8joo/MVTCAE

[5] https://github.com/miguelsvasco/nexus_pytorch

C.3 MLD with powerfull autoencoder

Here we provide more detail about the CUB experiment using more powerful autoencoder denoted MLD* in Figure 2. We use an architecture similar to [26] adapted to (64X64) resolution images. We modified the autoencoder architecture to be deterministic and train the model with a simple Mean square error loss. We kept the same configuration of the CUB experiment described in the previous experiment on the same dataset including the text autoencoder, score network and hyper-parameters.

C.4 Computation Resources

In our experiments, we used 4 A100 GPUs, for a total of roughly 4 months of experiments.

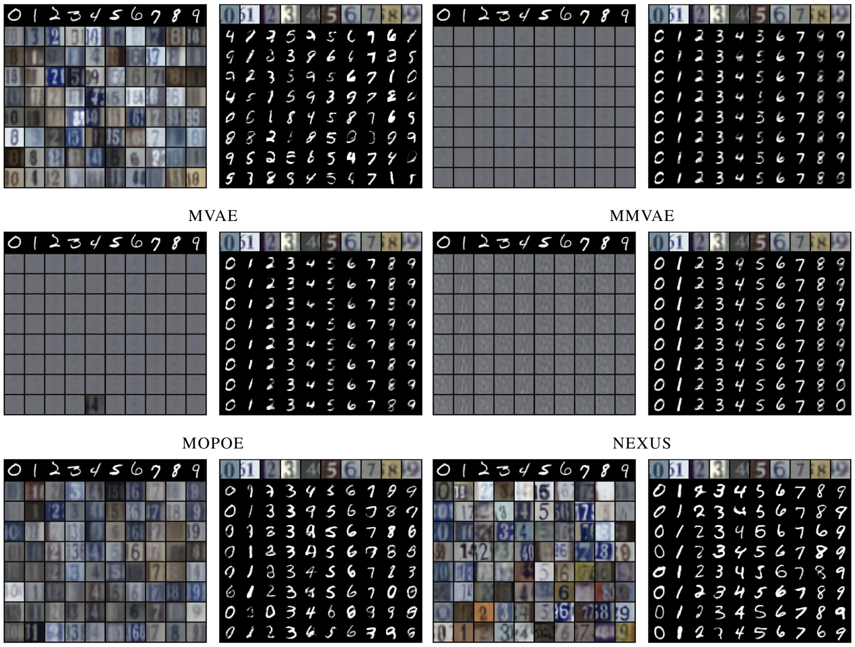
D Additional results

In this section, we report detailed results for all of our experiments, including standard deviation and additional qualitative samples, for all the data-sets and all the methods we compared in our work.

D.1 MNIST-SVHN

Table 4: Generative Coherence for **MNIST-SVHN**. We report the detailed version of Table 1 with standard deviation for 5 independent runs with different seeds.

Models	Coherence (%↑)			Quality (↓)			
	Joint	M → S	S → M	Joint(M)	Joint(S)	M → S	S → M
MVAE	$38.19_{\pm 2.27}$	$48.21_{\pm 2.56}$	$28.57_{\pm 1.46}$	$13.34_{\pm 0.93}$	$68.0_{\pm 0.99}$	$68.9_{\pm 1.84}$	$13.66_{\pm 0.95}$
MMVAE	$37.82_{\pm 1.19}$	$11.72_{\pm 0.33}$	$67.55_{\pm 9.22}$	$25.89_{\pm 0.46}$	$146.82_{\pm 4.76}$	$393.33_{\pm 4.86}$	$53.37_{\pm 1.87}$
MOPOE	$39.93_{\pm 1.54}$	$12.27_{\pm 0.68}$	$68.82_{\pm 0.39}$	$20.11_{\pm 0.96}$	$129.2_{\pm 6.33}$	$373.73_{\pm 26.42}$	$43.34_{\pm 1.72}$
NEXUS	$40.0_{\pm 2.74}$	$16.68_{\pm 5.93}$	$70.67_{\pm 0.77}$	$13.84_{\pm 1.41}$	$98.13_{\pm 5.9}$	$281.28_{\pm 16.07}$	$53.41_{\pm 1.54}$
MVTCAE	$48.78_{\pm 1}$	$\underline{81.97}_{\pm 0.32}$	$49.78_{\pm 0.88}$	$12.98_{\pm 0.68}$	$\mathbf{52.92}_{\pm 1.39}$	$69.48_{\pm 1.64}$	$13.55_{\pm 0.8}$
MLD	$\mathbf{85.22}_{\pm 0.5}$	$\mathbf{83.79}_{\pm 0.62}$	$\mathbf{79.13}_{\pm 0.38}$	$\mathbf{3.93}_{\pm 0.12}$	$\underline{56.36}_{\pm 1.63}$	$\mathbf{57.2}_{\pm 1.47}$	$\mathbf{3.67}_{\pm 0.14}$



MVAE

MMVAE

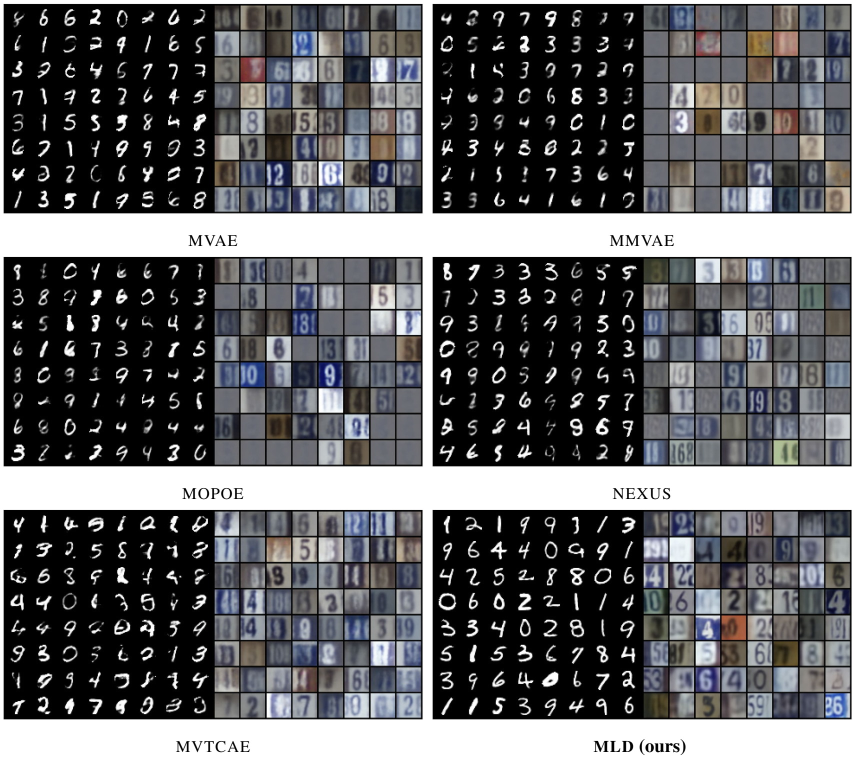
MOPOE

NEXUS

MVTCAE

MLD (ours)

Figure 7: Additional qualitative results for **MNIST-SVHN**. For each model we report: MNIST to SVHN conditional generation in the left, SVHN to MNIST conditional generation in the right.



MVAE

MMVAE

MOPOE

NEXUS

MVTCAE

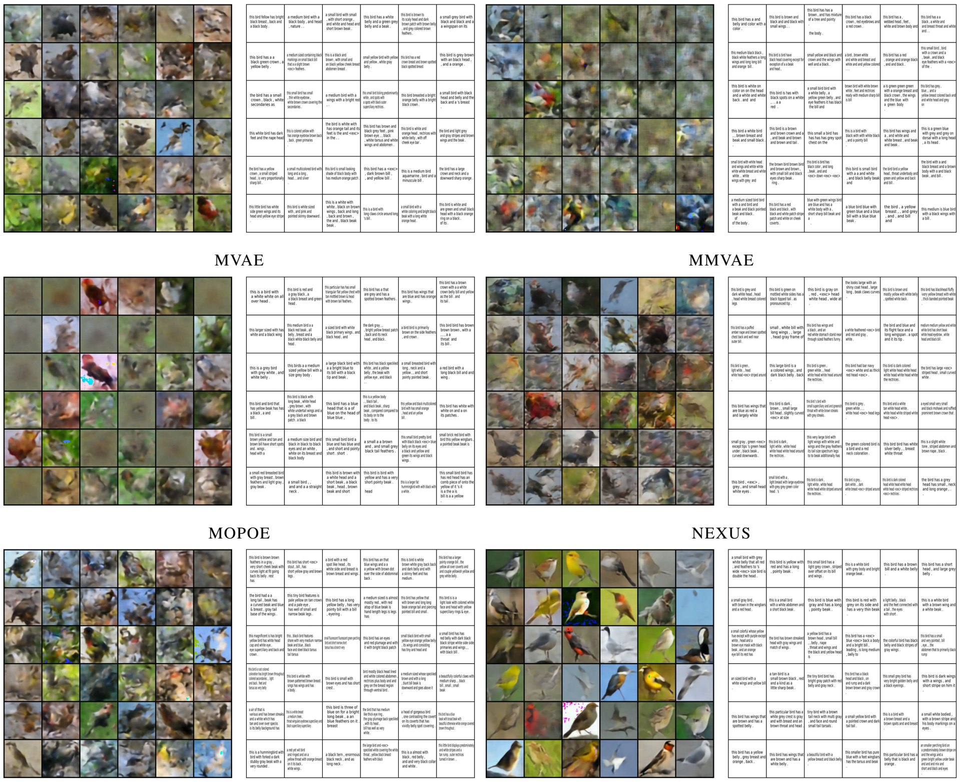
MLD (ours)

Figure 8: Qualitative results for **MNIST-SVHN** joint generation.

D.2 CUB

Models	Coherence (↑)			Quality (↓)	
	Joint	Image → Caption	Caption → Image	Joint → Image	Caption → Image
MVAE	0.66	**0.70**	0.64	158.91	158.88
MMVAE	0.66	0.69	0.62	277.8	212.57
MOPOE	0.64	0.68	0.55	279.78	179.04
NEXUS	0.65	0.69	0.59	147.96	262.9
MVTCAE	0.65	**0.70**	0.65	155.75	168.17
MLD	**0.69**	0.69	**0.69**	**63.47**	62.62
MLD*	**0.70**	0.69	**0.69**	**22.19**	**22.50**

Table 5: Generation Coherence (CLIP-S : Higher is better) and Quality (FID ↓ Lower is better) for CUB dataset. **MLD*** denotes the version of our method using a more powerful image autoencoder.



MVAE

MMVAE

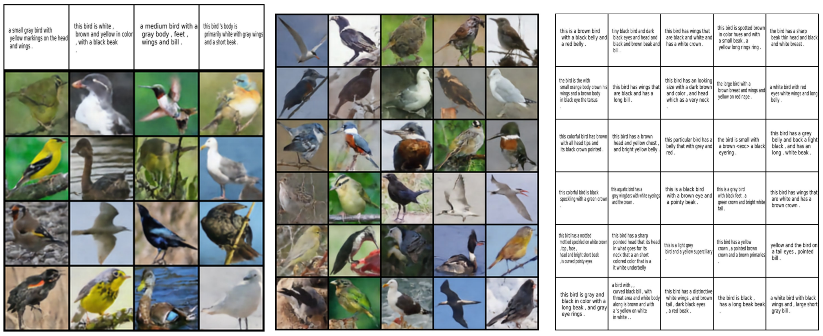
MOPOE

NEXUS

MVTCAE

MLD (ours)

Figure 9: Qualitative results for joint generation on **CUB**.


a small gray bird with yellow markings on the head and wings .
this bird is white , brown and yellow in color , with a black beak .
a medium bird with a gray body , feet , wings and bill .
this bird 's body is primarily white with gray wings and a short beak .
this is a brown bird with a black belly and a red belly .
tiny black bird and dark black eyes and head and black and brown beak and bill .
this bird has wings that are black and white and has a white crown .
this bird is spotted brown in color hues and with a small beak , a yellow long rings ring .
the bird has a sharp beak thin head and black and white breast .
the bird is the with small orange body crown his wings and a brown body in black eye the tarsus .
this bird has wings that are black and has a long bill .
this bird has an looking size with a dark brown and color , and head which as a very neck .
the large bird with a brown breast and wings and yellow on red nape .
a white bird with red eyes white wings and long belly .
this colorful bird has brown with all head tips and its black crown pointed .
this bird has a brown head and yellow chest , and bright yellow belly .
this particular bird has a belly that with grey and red .
the bird is small with a brown <exc> a black eyering .
this bird has a grey belly and back a light black , and has an long , white beak .
this colorful bird is black speckling with a green crown .
this aquatic bird has a grey wingbars with white eyerings and the crown .
this is a black bird with a brown eye and a pointy beak .
this is a gray bird with black feet , a green crown and bright white tail .
this bird has wings that are white and has a brown crown .
this bird has a mottled mottled speckled on white crown , top , face , head and bright short beak , is curved pointy eyes
this bird has a sharp pointed head that its head in what goes for its neck that a an short colored color that is a it white underbelly
this is a light grey bird and a yellow superciliary .
this bird has a yellow crown , a pointed brown crown and a brown primaries .
yellow and the bird on a tail eyes , pointed bill .
this bird is gray and black in color with a long beak , and gray eye rings .
a bird with , , curved black bill , with throat area and white body along is brown and with a 's yellow on white in white . .
this bird has a distinctive white wings , and brown tail , dark black eyes , a red beak .
the bird is black , has a long beak beak .
a white bird with black wings and , large short gray bill .

(a) Conditional generation.

(b) Joint generation.

Figure 10: Qualitative results of **MLD*** on **CUB** data-set with powerful image autoencoder.