Masked Multi-time Diffusion for Multi-modal Generative Modeling

Mustapha Bounoua[12], Giulio Franzese[2], Pietro Michiardi[2]

[1]Renault Software Factory, [2]EURECOM

Multi-modal Generative models

- Multi-modal generative models aim at approximating the distribution of multi-modal data (such as images, text, audio) while providing the capability to draw new samples either unconditionally (joint generation) or conditionally on a set of available modalities. These models are evaluated in terms of :
  - ***Quality*** of the generation which reflects the fidelity of the generated samples to the observed data.
  - ***Coherence*** of the generation in terms of consistency of the semantic information across the modalities.
- VAE-based multi-modal models have dominated this field, so far.

Limitations of multi-modal VAEs

- **Coherence-Quality trade-off.**
- **Latent collapse** impacting the quality of the latent variables.
- **Modality collapse** gradient-conflict.

Despite several efforts appearing in the recent literature, these limitations are **still not resolved** [1].

Contributions

Multi-modal Latent Diffusion (MLD) is a novel method for multi-modal generative modeling that, by design, does not suffer from the aforementioned limitations. Our approach is a two-stage procedure :

- Deterministic uni-modal autoencoders map the multi-modal data to the latent space. Their independent training avoids any gradient conflict.
- A score-based diffusion model is applied on the multi-modal latent space. The **multi-time masked diffusion process** enables joint and conditional generation for **any subset** of modalities.

Our approach achieves both ***high-quality and coherent*** joint/conditional data generation.

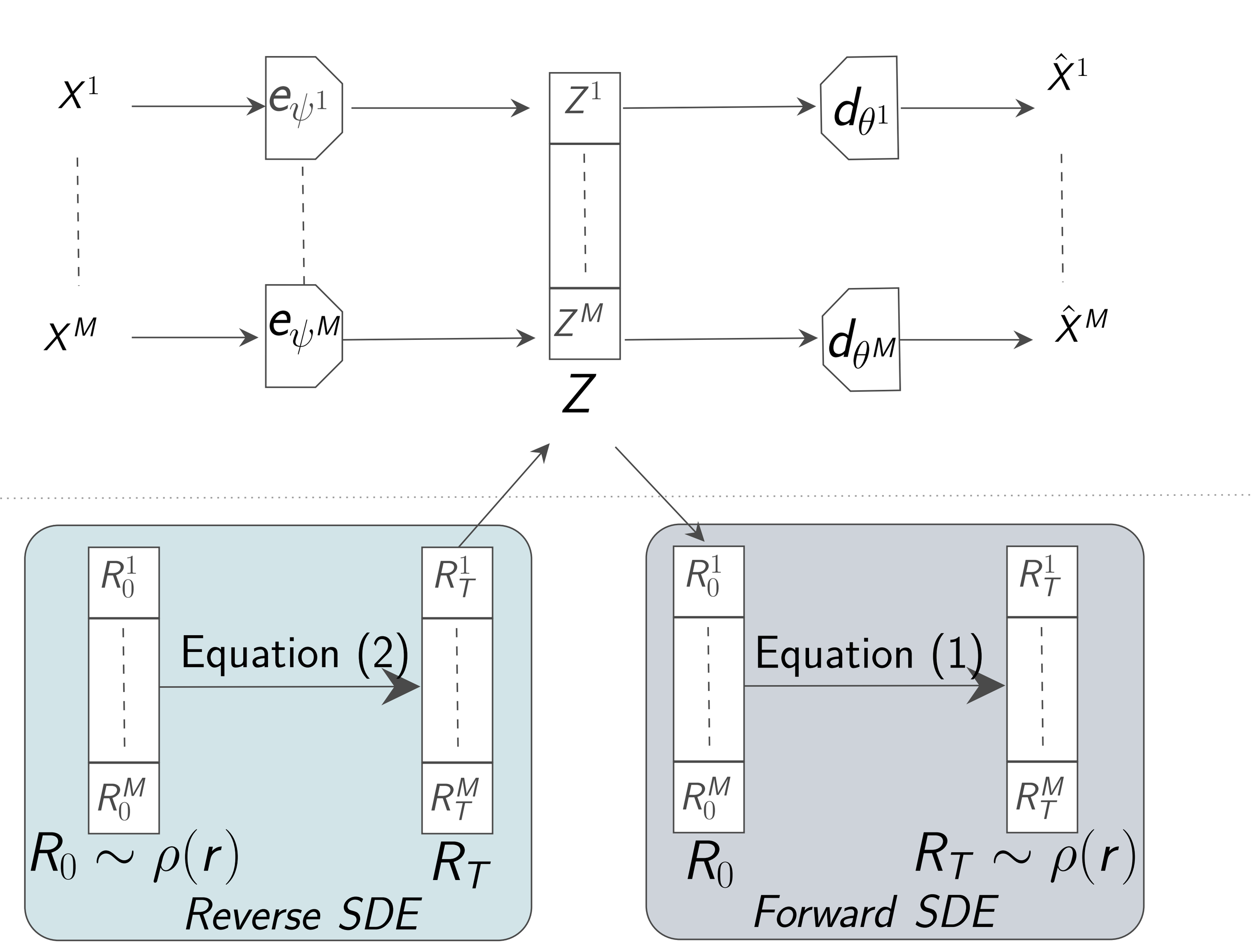

Figure: A schematic representation of **MLD**: ***Top:*** deterministic, modality-specific encoder/decoders, ***Bottom:*** A score-based diffusion model applied on the multi-modal latent space.

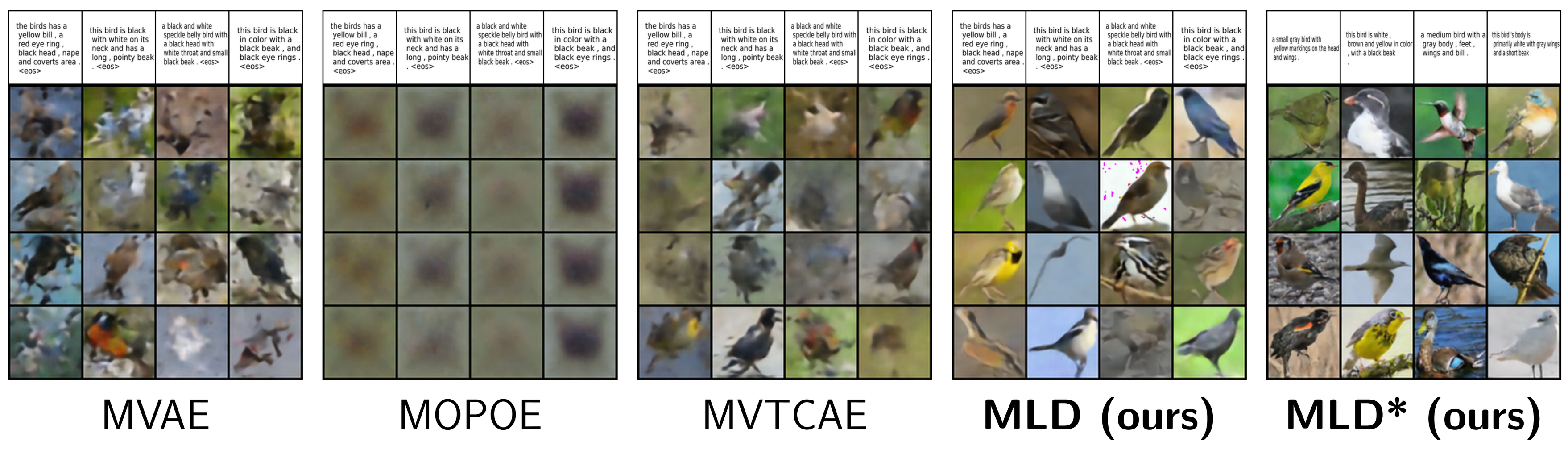

Figure: Qualitative results on CUB data-set. Caption used as condition to generate images. (**MLD***: denotes a version of our method using a powerful image autoencoder.)

Multi-modal Latent Diffusion

Given a generic partition of all modalities into non overlapping sets $A_1 \cup A_2$, we define the masked forward diffusion process.

The masked forward SDE

$$\mathrm{d}R_t = m(A_1) \odot [\alpha(t)R_t\mathrm{d}t + g(t)\mathrm{d}W_t], \tag{1}$$

where $\alpha(t)R_t$ and $g(t)$ are the drift and diffusion terms, respectively, and W_t is a Wiener process. The mask $m(A_1)$ ensures that only the portion of the latent space concerning the modalities of the subset A_1 is diffused.

To sample from $q_\psi(z^{A_1} \mid z^{A_2})$, we derive the reverse-time dynamics of eq. (1):

The masked reverse SDE

$$\mathrm{d}R_t = m(A_1) \odot \left[\left(-\alpha(t')R_t + g^2(t')\nabla \log\left(q(R_t, t' \mid z^{A_2})\right)\right)\mathrm{d}t + g(t')\mathrm{d}W_t\right], \tag{2}$$

with $t' = T - t$, the initial conditions $R_0 = \mathcal{C}(R_0^{A_1}, z^{A_2})$ and $R_0^{A_1} \sim \rho(r^{A_1})$. The true score function $\nabla \log\left(q(r, t \mid z^{A_2})\right)$ is approximated by a conditional score network $s_\chi(r^{A_1}, t \mid z^{A_2})$.

Multi-time Diffusion

Instead of training a separate score network for each possible combination, we use a single architecture that accepts all modalities as input and a *multi-time vector* $\tau = [t_1, \dots, t_M]$. The multi-time vector serves as a conditioning signal and additionally indicates the diffusion time.

- **Training:** At each step, a randomly selected set of modalities A_1 is diffused while A_2 is freezed during the forward process.
- **Generation:** Any valid numerical integration scheme (E.g. Euler-Maruyama) for eq. (2) can be used for conditional generation.

MLD captures efficiently the interactions across modalities

MLD treats the multi-modal latent space as variables that evolve differently through the diffusion process according to a multi-time vector. Each modality diffusion time modulates its influence on the generation.

- Joint generation (same diffusion time) : All the modalities influence each other equally.
- Conditional generation: The conditioning modalities are not perturbed which reflects a maximal influence on the generation.

Table: Generation coherence and quality for **MNIST-SVHN** (M :MNIST, S: SVHN). *Quality* is measured in terms of FID. *Coherence* is measured as in [3, 4, 2], using pre-trained classifiers.

Models	Coherence (%↑)			Quality (↓)			
	Joint	M → S	S → M	Joint(M)	Joint(S)	M → S	S → M
MVAE [6]	38.19	48.21	28.57	13.34	68.9	<u>68.0</u>	13.66
MMVAE [3]	37.82	11.72	67.55	25.89	146.82	393.33	53.37
MOPOE [4]	39.93	12.27	68.82	20.11	129.2	373.73	43.34
NEXUS [5]	40.0	16.68	<u>70.67</u>	13.84	98.13	281.28	53.41
MVTCAE [2]	<u>48.78</u>	<u>81.97</u>	49.78	<u>12.98</u>	**52.92**	69.48	<u>13.55</u>
MLD	**85.22**	**83.79**	**79.13**	**3.93**	<u>56.36</u>	**57.2**	**3.67**

References

[1] *Daunhawer Imant, Sutter Thomas M., Chin-Cheong Kieran, Palumbo Emanuele, Vogt Julia E.* On the Limitations of Multimodal VAEs // International Conference on Learning Representations. 2022.

[2] *Hwang HyeongJoo, Kim Geon-Hyeong, Hong Seunghoon, Kim Kee-Eung.* Multi-View Representation Learning via Total Correlation Objective // Advances in Neural Information Processing Systems. 2021. 34. 12194–12207.

[3] *Shi Yuge, N Siddharth, Paige Brooks, Torr Philip.* Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models // Advances in Neural Information Processing Systems. 32. 2019.

[4] *Sutter Thomas M., Daunhawer Imant, Vogt Julia E.* Generalized Multimodal ELBO // International Conference on Learning Representations. 2021.

[5] *Vasco Miguel, Yin Hang, Melo Francisco S., Paiva Ana.* Leveraging hierarchy in multimodal generative models for effective cross-modality inference // Neural Networks. 2 2022. 146. 238–255.

[6] *Wu Mike, Goodman Noah.* Multimodal Generative Models for Scalable Weakly-Supervised Learning // Advances in Neural Information Processing Systems. 31. 2018.