entropy

MDPI

Article

Multi-Modal Latent Diffusion

Mustapha Bounoua [1,2,*], Giulio Franzese [2] and Pietro Michiardi [2]

[1] Ampere Software Technology, 06560 Valbonne, France
[2] Department of Data Science, EURECOM, 06410 Biot, France; giulio.franzese@eurecom.fr (G.F.); pietro.michiardi@eurecom.fr (P.M.)
* Correspondence: mustapha.bounoua@eurecom.fr

Abstract: Multimodal datasets are ubiquitous in modern applications, and multimodal Variational Autoencoders are a popular family of models that aim to learn a joint representation of different modalities. However, existing approaches suffer from a coherence–quality tradeoff in which models with good generation quality lack generative coherence across modalities and vice versa. In this paper, we discuss the limitations underlying the unsatisfactory performance of existing methods in order to motivate the need for a different approach. We propose a novel method that uses a set of independently trained and unimodal deterministic autoencoders. Individual latent variables are concatenated into a common latent space, which is then fed to a masked diffusion model to enable generative modeling. We introduce a new multi-time training method to learn the conditional score network for multimodal diffusion. Our methodology substantially outperforms competitors in both generation quality and coherence, as shown through an extensive experimental campaign.

Keywords: multimodality; generative models; score-based models; diffusion models

Citation: Bounoua, M.; Franzese, G.; Michiardi, P. Multi-Modal Latent Diffusion. *Entropy* **2024**, *26*, 320. https://doi.org/10.3390/e26040320

Academic Editors: Sotiris Kotsiantis and Jakub Tomczak

Received: 12 February 2024
Revised: 29 March 2024
Accepted: 31 March 2024
Published: 5 April 2024

1. Introduction

Multi-modal generative modeling is a crucial area of research in machine learning that aims to develop models capable of generating data according to multiple modalities, such as images, text, audio, and more. This is important because real-world observations are often captured in various forms; thus, 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, while audio and video can capture different aspects of a scene. Multimodal generative models can help in tasks such as data augmentation [1–3], missing modality imputation [4–7], and conditional generation [8,9].

Multimodal models have flourished over the past years and seen tremendous interest from academia and industry, especially in the content creation sector. Whereas most recent approaches focus on specialization, by considering text as a primary input to be associated mainly with images [10–16] and videos [17–19], in this work we target an established literature with more general scope and in which all modalities are considered equally important.

Multi modal generative models aim at *high-quality* data generation, as well as at 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. The predominant literature in this field is based on extensions of the Variational Autoencoder (VAE) [20] to the multimodal domain; initially interested in learning joint latent representation of multimodal data, such works have mostly focused on generative modeling.

In short, multimodal VAEs relies on combinations of unimodal VAEs, and the design space mainly consists of the way in which the unimodal latent variables are combined to construct the joint posterior distribution. Early works such as [21] adopted

a product-of-experts approach, whereas others [22] considered a mixture-of-experts approach. While product-based models achieve high generative quality, they suffer in terms of both joint and conditional coherence. This has been found to be due to mis-calibration issues on the part of the experts [22,23]. On the other hand, mixture-based models produce coherent but qualitatively poor samples. A first attempt to address the so-called **coherence–quality tradeoff** [24] was represented by the mixture of products of experts approach [23]. However, recent comparative studies [24] have shown that none of the existing approaches fulfill the criteria of both generative quality and coherence. A variety of techniques are aimed at finding a better operating point, such as contrastive learning techniques [25], hierarchical schemes [26], total correlation-based calibration of single-modality encoders [27], and different training objectives [28]. More recently, in [29], explicitly separated shared and private latent spaces were considered as a way to overcome the aforementioned limitations.

In Section 2, we investigate the limitations of multimodal VAEs and prepared the ground to substantiate a new approach which overcomes the shortcomings in the state of the art. We further investigate the tradeoff [24] between generative coherence and quality, and argue that it is intrinsic to all variants of multimodal VAEs. We indicate two root causes of the problem: latent variable collapse [30,31] and information loss due to mixture subsampling. To tackle these issues, in Section 3 of this work we propose a new approach that uses a set of independent and unimodal *deterministic* autoencoders with the latent variables simply concatenated in a joint latent variable. Joint and conditional generative capabilities are provided by an additional model that learns a probability density associated with the joint latent variable. We propose an extension of score-based diffusion models [32] to operate on the multimodal latent space. Thus, we derive both forward and backward dynamics that are compatible with the multimodal nature of the latent data. In Section 4, we propose a novel multi-time diffusion process that can both be used for joint and conditional generation. We label our approach Multi-modal Latent Diffusion (MLD).

Our experimental evaluation of MLD in Section 5 provides compelling evidence of the superiority of our approach for multimodal generative modeling. We compare MLD to a large variety of VAE-based alternatives on several real-life multimodal datasets in terms of generative quality and both joint and conditional coherence. Our model outperforms alternatives in all possible scenarios, even those that are notoriously difficult because the modalities might be only loosely correlated. We note that recent works have explored the joint generation of multiple modalities [33,34]; however, such approaches are application-specific, e.g., text-to-image, and essentially only target two modalities. When relevant, we compare our method to additional recent alternatives to multimodal diffusion [35,36] and show the superior performance of MLD.

2. Limitations of Multimodal VAEs

In this work, we consider multimodal VAEs [21–23,29] as the standard modeling approach to tackle both joint and conditional generation of multiple modalities. Our goal here is the need to go beyond such a standard approach in order to overcome limitations that affect multimodal VAEs, which result in a tradeoff between generation quality and generative coherence [24,29].

Consider the random variable $X = \{X^1, \ldots, X^M\} \sim p_D(x^1, \ldots, x^M)$, consisting of the set M of modalities sampled from the (unknown) multimodal data distribution p_D. We indicate the marginal distribution of a single modality by $X^i \sim p_D^i(x^i)$ and the collection of a generic subset of modalities by $X^A \sim p_D^A(x^A)$, with $X^A \stackrel{\text{def}}{=} \{X^i\}_{i \in A}$, where $A \subset \{1, \ldots, M\}$ is a set of indexes; for example, given $A = \{1, 3, 5\}$, we would have $X^A = \{X^1, X^3, X^5\}$.

We begin by considering unimodal VAEs as particular instances of the Markov chain $X \to Z \to \hat{X}$, where Z is a latent variable and $\hat{X}$ is the generated variable. Models are specified by the two conditional distributions, called the encoder $Z \,|_{X=x} \sim q_\psi(z \,|\, x)$ and decoder $\hat{X} \,|_{Z=z} \sim p_\theta(\hat{x} \,|\, z)$. For a given prior distribution $p_n(z)$, the objective is to define a generative model with samples that are distributed as similarly as possible to the original data.

In the case of multimodal VAEs, we consider the general family of Mixture of Product of Experts (MOPOE) [23], which includes as particular cases many existing variants such as Product of Experts (MVAE) [21] and Mixture of Expert (MMVAE) [22]. Formally, a collection of K arbitrary subsets of modalities $S = \{A_1, \dots A_K\}$ along with weighting coefficients $\omega_i \geq 0, \sum_{i=1}^{K} \omega_i = 1$ define the posterior $q_\psi(z \mid x) = \sum_i \omega_i q^i_{\psi^{A_i}}(z \mid x^{A_i})$, with $\psi = \{\psi^1, \dots, \psi^K\}$. To lighten the notation, we use $q_{\psi^{A_i}}$ in place of $q^i_{\psi^{A_i}}$, noting that the various $q^i_{\psi^{A_i}}$ can have both different parameters ψ^{A_i} and functional forms. For example, in the MOPOE [23] parametrization, we have $q_{\psi^{A_i}}(z \mid x^{A_i}) = \prod_{j \in A_i} q_{\psi^j}(z \mid x^j)$. Our exposition is more general, and is not limited to this assumption. The selection of the posterior can be understood as the result induced by the two step procedure where (i) each subset of modalities A_i is encoded into specific latent variables $Y_i \sim q_{\psi^{A_i}}(\cdot \mid x^{A_i})$ and (ii) the latent variable Z is obtained as $Z = Y_i$ with probability ω_i. Optimization is performed with respect to the following evidence lower bound (ELBO) [23,24]:

$$\mathcal{L} = \sum_i \omega_i \int p_D(x) q_{\psi^{A_i}}(z \mid x^{A_i}) \log p_\theta(x|z) - \log \frac{q_{\psi^{A_i}}(z \mid x^{A_i})}{p_n(z)} \mathrm{d}z\mathrm{d}x. \tag{1}$$

A well known limitation called the latent collapse problem [30,31] affects the quality of the latent variables Z. Consider the hypothetical case of arbitrary flexible encoders and decoders. Posteriors with zero mutual information with respect to model inputs are valid maximizers of Equation (1). To prove this, it is sufficient to substitute the posteriors $q_{\psi^{A_i}}(z \mid x^{A_i}) = p_n(z)$ and $p_\theta(x|z) = p_D(x)$ into Equation (1) to observe that the optimal value of $\mathcal{L} = \int p_D(x) \log p_D(x) \mathrm{d}x$ is achieved [30,31]. The problem of information loss is exacerbated in the case of multimodal VAEs [24]. Intuitively, even if the encoders $q_{\psi^{A_i}}(z \mid x^{A_i})$ carry relevant information about their inputs X^{A_i}, step (ii) of the multimodal encoding procedure described above induces a further information bottleneck. Some fraction ω_i of the time, the latent variable Z will be a copy of Y_i, which only provides information about the subset X^{A_i}. No matter how good the encoding step is, the information about $X^{\{1,\dots,M\} \setminus A}$ that is not contained in X^{A_i} cannot be retrieved.

The variable collapse problem can be analyzed through the lenses of self-reconstruction, whereby a multimodal VAE is evaluated by simply reconstructing the same modality it receives as input. We have observed that these models tend to encode input samples into a latent space with possible information loss, leading to inconsistent reconstruction. This is particularly shown by the quantitative results in Table A7, with notable difficulty in reconstructing the SVHN modality.

Furthermore, if the latent variable carries zero mutual information with respect to the multimodal input, a coherent *conditional* generation of a set of modalities given others is impossible, as $\hat{X}^{A_1} \perp X^{A_2}$ for any generic sets A_1, A_2. While the factorization $p_\theta(x \mid z) = \prod_{i=1}^{M} p_{\theta^i}(x^i \mid z), \theta = \{\theta_1, \dots, \theta_M\}$ (we use p_{θ^i} here instead of $p^i_{\theta^i}$ to unclutter the notation) could enforce preservation of information and guarantee better quality of the *jointly* generated data, in practice the latent collapse phenomenon induces multimodal VAEs to converge towards suboptimal a operating regime. When the posterior $q_\psi(z \mid x)$ collapses onto the uninformative prior $p_n(z)$, the ELBO in Equation (1) reduces to the sum of modality-independent reconstruction terms:

$$\sum_i \omega_i \sum_{j \in A_i} \int p^j_D(x^j) p_n(z) \Big(\log p_{\theta^j}(x^j|z) \Big) \mathrm{d}z\mathrm{d}x^j \tag{2}$$

where, paradoxically, the quality of the approximation of the various marginal distributions is extremely high, while there is a complete lack of joint coherence.

General principles to avoid latent collapse involve explicitly forcing the learning of informative encoders $q_\theta(z \mid x)$ via $\beta-$annealing of the Kullback-Leibler (KL) term in the ELBO and the reduction of the representational power of encoders and decoders.

While $\beta-$annealing [37] has been explored in the multimodal VAEs literature, [21] with limited improvements reported, reducing the flexibility of the encoders/decoders clearly impacts the generation quality. Hence, the presence of the tradeoff; in order to improve coherence, the flexibility of encoders/decoders should be constrained, which in turn impacts generative quality. This tradeoff has recently been addressed in the literature on multimodal VAEs [24,29]; however, our experimental results in Section 5 indicate that there is ample room for improvement and that a new approach is truly needed.

3. Our Approach: Multimodal Latent Diffusion

We propose a new method for multimodal generative modeling that by design does not suffer from the limitations discussed in Section 2. Our objective is to enable both high quality and coherent joint/conditional data generation using a simple design (see Figure 1 for a schematic representation). As an overview, we use deterministic unimodal autoencoders whereby each modality X^i is encoded through its encoder e_{ψ^i} (which is a short form for $e^i_{\psi^i}$) into the modality-specific latent variable Z^i and decoded into the corresponding $\hat{X}^i = d_{\theta^i}(Z^i)$. Our approach can be interpreted as a latent variable model in which the different latent variables Z^i are concatenated as $Z = [Z^1, \ldots, Z^M]$. This corresponds to the parameterization of the two conditional distributions as $q_\psi(z \mid x) = \prod_{i=1}^{M} \delta(z^i - e_{\psi^i}(x^i))$ and $p_\theta(\hat{x} \mid z) = \prod_{i=1}^{M} \delta(\hat{x}^i - d_{\theta^i}(z^i))$, respectively. Then, in place of an ELBO, we optimize the parameters of our autoencoders by minimizing the following sum of modality-specific losses:

$$\mathcal{L} = \sum_{i=1}^{M} \mathcal{L}_i, \quad \mathcal{L}_i = \int p^i_D(x^i) l^i(x^i - d_{\theta^i}(e_{\psi^i}(x^i))) \mathrm{d}x^i, \tag{3}$$

where l^i can be any valid distance function, e.g, the square norm $\|\cdot\|^2$. The parameters ψ^i, θ^i are modality-specific; thus, minimization of Equation (3) corresponds to individual training of the different autoencoders. Because the mapping from input to latent is deterministic, there is no loss of information between X and Z (note that as the measures are not absolutely continuous with respect to the Lebesgue measure, the mutual information is $+\infty$). Moreover, this choice avoids any form of interference in the backpropagated gradients corresponding to the unimodal reconstruction losses. Consequently, gradient conflict issues [38], in which stronger modalities pollute weaker ones, are avoided.

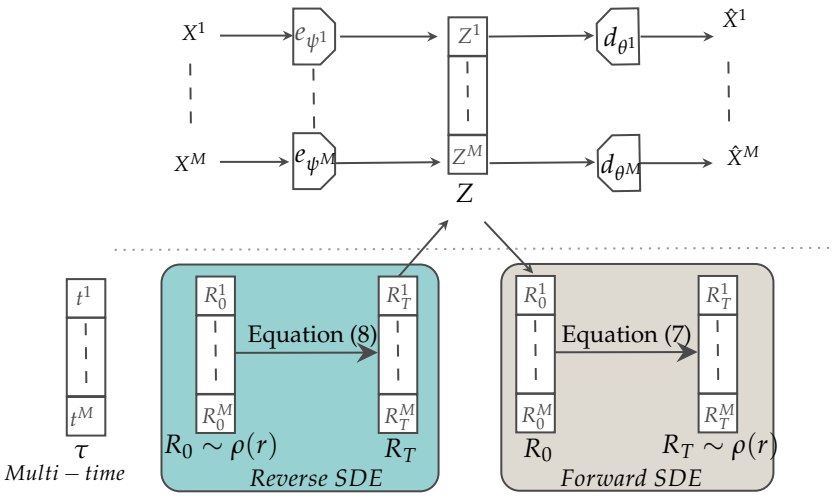


Figure 1. Multimodal Latent Diffusion: two-stage model involving (**Top**): deterministic modality-specific encoder/decoders and (**Bottom**): the score-based diffusion model on the latent spaces of the modalities, which evolve differently through the diffusion process according to a multi-time vector.

To enable such a simple design to become a generative model, it is sufficient to generate samples from the induced latent distribution $Z \sim q_\psi(z) = \int p_D(x) q_\psi(z \mid x) \mathrm{d}x$ and decode them as $\hat{X} = d_\theta(Z) = [d_{\theta^1}(Z^1), \ldots, d_{\theta^M}(Z^M)]$.

To obtain such samples, we follow the two-stage procedure described in [39–41], where samples from the lower-dimensional $q_\psi(z)$ are obtained through a score-based generative model. These models have shown tremendous performance in fitting complex distributions [10,42], an ability which aligns with our objective of learning the distribution within a multimodal latent space. Furthermore, the conditioning mechanism inherent in score-based models facilitates highly coherent generation. MLD is further enhanced by a multi-time diffusion process, a novel mechanism that allows for the generation of any subset of modalities, and which we explain in Section 4.

It may be helpful at this point to clarify that the two-stage training of MLD is carried out separately. Unimodal deterministic autoencoders are pretrained first, followed by the training of the score-based diffusion model, which is explained in more detail later.

To conclude this overview of our method, for joint data generation it is possible to sample from noise, perform backward diffusion, and then decode the generated multimodal latent variable to obtain the corresponding data samples. For conditional data generation, given one modality, the reverse diffusion is guided by this modality, while the other modalities are generated by sampling from noise. The generated latent variable is then decoded to obtain data samples of the missing modality.

Joint and Conditional Multimodal Latent Diffusion Processes

In the first stage of our method, 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 $Z = [Z^1, \ldots, Z^M]$, resulting from the concatenation of unimodal latent variables.

Joint generation: To generate a new sample for all modalities, we use a simple score-based diffusion model in latent space [32,39,40,42,43]. 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

$$\mathrm{d}R_t = \alpha(t) R_t \mathrm{d}t + g(t) \mathrm{d}W_t, \quad R_0 \sim q(r, 0), \tag{4}$$

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 [44] with initial conditions $q(r, 0)$. We assume the uniqueness and existence of a stationary distribution $\rho(r)$ for the process in Equation (4), though this is not necessary for the validity of the method [45]. 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 [46], a time-reversed stochastic process exists, with a new SDE of the form

$$\mathrm{d}R_t = \Big(-\alpha(T-t) R_t + g^2(T-t) \nabla \log(q(R_t, T-t))\Big) \mathrm{d}t + g(T-t) \mathrm{d}W_t \quad R_0 \sim q(r, T), \tag{5}$$

indicating that, in principle, simulation of Equation (5) allows samples to be generated from the desired distribution $q(r, 0)$. In practice, we use a **parametric score network** $s_\chi(r, t)$ to approximate the true score function, and we approximate $q(r, T)$ with the stationary distribution $\rho(r)$. Indeed, the generated data distribution $q(r, 0)$ is close (in the KL sense) to the true density as described by [45,47]:

$$\mathrm{KL}[q_\psi(r) \mid\mid q(r, 0)] \leq \frac{1}{2} \int_0^T g^2(t) \mathbb{E}[\| s_\chi(R_t, t) - \nabla \log q(R_t, t) \|^2] \mathrm{d}t + KL[q(r, T) || \rho(r)] \tag{6}$$

where the first term on the right-hand side is referred to as the score-matching objective, and is the loss over which the score network is optimized, while the second is a vanishing term for $T \to \infty$.

To conclude, joint generation of all modalities is achieved through simulation of the reverse-time SDE in Equation (5), followed by a simple decoding procedure. Indeed, optimally trained decoders (achieving zero in Equation (3)) can be used to transform $Z \sim q_\psi(z)$ into samples from $\int p_\theta(x \mid z) q_\psi(z) \mathrm{d}z = p_D(x)$.

Conditional generation. Given a generic partition of all modalities into non-overlapping sets $A_1 \cup A_2$, where $A_2 = (\{1, \ldots, M\} \setminus A_1)$, 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 $R_0 = \mathcal{C}(R_0^{A_1}, R_0^{A_2})$ and 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}). As an illustration, consider $A_1 = \{1, 3, 5\}$ such that $X^{A_1} = \{X^1, X^3, X^5\}$ and $A_2 = \{2, 4, 6\}$ such that $X^{A_2} = \{X^2, X^4, X^6\}$; then, $R_0 = \mathcal{C}(R_0^{A_1}, R^{A_2}) = \mathcal{C}(R_0^{A_1}, z^{A_2}) = [R_0^1, z^2, R_0^3, z^4, R_0^5, z^6]$.

More formally, we define the following masked forward-diffusion SDE:

$$\mathrm{d}R_t = m(A_1) \odot [\alpha(t) R_t \mathrm{d}t + g(t) \mathrm{d}W_t], \;\; q(r, 0) = q_\psi(r^{A_1} \mid z^{A_2}) \delta(r^{A_2} - z^{A_2}) \tag{7}$$

The mask $m(A_1)$ contains M vectors u^i, one per modality, with the corresponding cardinality. If modality $j \in A_1$, then $u^j = 1$; otherwise, $u^j = 0$. Then, the effect of masking is to "freeze" the part of the random variable R_t corresponding to the conditioning latent modalities z^{A_2} throughout the diffusion process. We naturally associate the conditional time-varying density $q(r, t \mid z^{A_2}) = q(r^{A_1}, t \mid z^{A_2}) \delta(r^{A_2} - z^{A_2})$ with this modified forward process.

To sample from $q_\psi(z^{A_1} \mid z^{A_2})$, we derive the reverse-time dynamics of Equation (7) as follows:

$$\mathrm{d}R_t = m(A_1) \odot \left[\left(-\alpha(T-t) R_t + g^2(T-t) \nabla \log \left(q(R_t, T-t \mid z^{A_2}) \right) \right) \mathrm{d}t + g(T-t) \mathrm{d}W_t \right] \tag{8}$$

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 $\nabla \log \left(q(r, t \mid z^{A_2}) \right)$ by a conditional score network $s_\chi(r^{A_1}, t \mid z^{A_2})$.

4. Multi-Time Diffusion to Learn the Conditional Score Network

A correctly optimized score network $s_\chi(r, t)$ allows samples from the joint distribution $q_\psi(z)$ to be obtained through simulation of Equation (5). Similarly, through the simulation of Equation (8), a *conditional* score network $s_\chi(r^{A_1}, t \mid z^{A_2})$ allows for sampling from $q_\psi(z^{A_1} \mid z^{A_2})$. In Section 4.1, we extend the guidance mechanisms used in classical diffusion models to allow multimodal conditional generation. A naïve alternative is to rely on the unconditional score network $s_\chi(r, t)$ for the conditional generation task by casting it as an *in-painting* objective. Intuitively, any missing modality could be recovered in the same way that a unimodal diffusion model can recover masked information. In Section 4.3, we discuss the implicit assumptions underlying in-painting from an information-theoretic perspective and argue that such assumptions are difficult to satisfy in the context of multimodal data. This intuition is corroborated by ample empirical evidence, where our method consistently outperforms alternatives.

4.1. Multi-Time Diffusion

We propose a modification to the classifier-free guidance technique [48] to learn a score network that can generate conditional and unconditional samples from any subset of modalities. 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* $\tau = [t_1, \ldots, t_M]$. 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 Section 3, we consider an arbitrary partitioning of all modalities in two disjoint sets, A_1 and A_2; set A_2 contains randomly selected conditioning modalities, while the remaining modalities belong to set A_1. During training, the parametric score network estimates $\nabla \log\big(q(r, t \,|\, z^{A_2})\big)$, whereby set A_2 is randomly chosen at every step. This is achieved by the *masked diffusion process* from Equation (7), 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 $\tau(A_1, t) = t\big[\mathbb{1}(1 \in A_1), \ldots, \mathbb{1}(M \in A_1)\big]$. As a follow-up of the example in Section 3, given $A_1 = \{1, 3, 5\}$ such that $X^{A_1} = \{X^1, X^3, X^5\}$ and $A_2 = \{2, 4, 6\}$ such that $X^{A_2} = \{X^2, X^4, X^6\}$, we have $\tau(A_1, t) = [t, 0, t, 0, t, 0]$.

More precisely, the algorithm for multi-time diffusion training (see Appendix A for the pseudo-code) proceeds as follows. At each step, a set of conditioning modalities A_2 is sampled from a predefined distribution ν, where $\nu(\emptyset) \stackrel{\text{def}}{=} \Pr(A_2 = \emptyset) = d$ and $\nu(U) \stackrel{\text{def}}{=} \Pr(A_2 = U) = {}^{(1-d)}/_{(2^M-1)}$ with $U \in \mathcal{P}(\{1, \ldots, M\}) \setminus \emptyset$, where $\mathcal{P}(\{1, \ldots, M\})$ is the powerset of all modalities. The corresponding set A_1 and mask $m(A_1)$ are constructed, and a sample X is drawn from the training dataset. The corresponding latent variables $Z^{A_1} = \{e_\psi^i(X^i)\}_{i \in A_1}$ and $Z^{A_2} = \{e_\psi^i(X^i)\}_{i \in A_2}$ are computed using the pretrained encoders and a diffusion process starting from $R_0 = \mathcal{C}(Z^{A_1}, Z^{A_2})$ is simulated for a randomly chosen diffusion time t using the conditional forward SDE with the mask $m(A_1)$. The score network is then fed the current state R_t and multi-time vector $\tau(A_1, t)$ and the difference between the score network's prediction and the true score is computed while applying mask $m(A_1)$. The score network parameters are updated using stochastic gradient descent, and this process is repeated for a total of L training steps. Clearly, when $A_2 = \emptyset$, training proceeds the same as for an unmasked diffusion process, as mask $m(A_1)$ allows all of the latent variables to be diffused.

Conditional generation: Any valid numerical integration scheme for Equation (8) can be used for conditional sampling (see Appendix A for an implementation using the Euler–Maruyama integrator). First, conditioning modalities in 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 a step size of $\Delta t = {}^T/_N$, starting from 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 as $\hat{X}^{A_1} = \{d_\theta^j(R_T^j)\}_{j \in A_1}$. Inference time conditional generation is not randomized; the conditioning modalities are the ones that are available, whereas those remaining are the ones we wish to generate.

Any-to-any multimodality has been recently studied through the composition of modality-specific diffusion models [49] by designing cross-attention and training procedures that allow for arbitrary conditional generation. This work by Tang et al. [49] relies on latent interpolation of input modalities, which is akin to mixture models, and uses it as conditioning signal for individual diffusion models. This is substantially different from the joint nature of the multimodal latent diffusion we present in our work; instead of forcing entanglement through cross-attention between score networks, our model relies on a joint diffusion process whereby modalities naturally co-evolve according. Another recent work [50], targeted multimodal conversational agents, wherein the strong underlying assumption is to consider one modality, i.e., text, as a guide for the alignment and generation of other modalities. Even if conversational objectives are orthogonal to our work, techniques akin to instruction-following for cross-generation are an interesting illustration of the powerful capabilities of in-context learning on the part of LLMs [51,52].

4.2. Multimodal Interaction

MLD treats the latent spaces of each modality as variables that evolve differently through the diffusion process according to a multi-time vector. The masked multi-time training enables the model to learn the score of all the combinations of conditionally diffused modalities, using the frozen modalities as the conditioning signal through a randomized scheme. By learning the score function of the diffused modalities at different time steps, the score model captures the correlation between the modalities.

At test time, the diffusion time of each modality is chosen so as to modulate its influence on the generation. For joint generation, the model uses the unconditional score, which corresponds to using the same diffusion time for all modalities. Thus, all the modalities influence each other equally. This ensures that the modality interaction information is faithful to the information characterizing the observed data distribution. The model can also generate modalities conditionally using the conditional score by freezing the conditioning modalities during the reverse process. The frozen state is similar to the final state of the revere process, where information is not perturbed; thus, the influence of the conditioning modalities is maximal. Subsequently, the generated modalities reflect the necessary information from the conditioning modalities and achieve the desired correlation.

4.3. In-Painting and Its Implicit Assumptions

Under certain assumptions, given an unconditional score network $s_\chi(r,t)$ that approximates the true score $\nabla \log q(r,t)$, it is possible to obtain a conditional score network $s_\chi(r^{A_1}, t \,|\, z^{A_2})$ to approximate $\nabla \log q(r^{A_1}, t \,|\, z^{A_2})$. We start by observing the equality

$$q(r^{A_1}, t \,|\, z^{A_2}) = \int q(\mathcal{C}(r^{A_1}, r^{A_2}), t \,|\, z^{A_2}) \mathrm{d}r^{A_2} = \int \frac{q(z^{A_2} \,|\, \mathcal{C}(r^{A_1}, r^{A_2}), t)}{q_\psi(z^{A_2})} q(\mathcal{C}(r^{A_1}, r^{A_2}), t) \mathrm{d}r^{A_2}, \tag{9}$$

where, with a slight abuse of notation, we indicate with $q(z^{A_2} \,|\, \mathcal{C}(r^{A_1}, r^{A_2}), t)$ the density associated with the event; the portion corresponding to A_2 of the latent variable Z is equal to z^{A_2}, given that the whole diffused latent R_t at time t is equal to $\mathcal{C}(r^{A_1}, r^{A_2})$. In the literature, the quantity $q(z^{A_2} \,|\, \mathcal{C}(r^{A_1}, r^{A_2}), t)$ is typically approximated by dropping its dependency on r^{A_1}. This approximation can be used to manipulate Equation (9) as $q(r^{A_1}, t \,|\, z^{A_2}) \simeq \int q(r^{A_2}, t \,|\, z^{A_2}) q(r^{A_1}, t | r^{A_2}, t) \mathrm{d}r$. Further, Monte Carlo approximations [32,53] of the integral allows for implementation of a practical scheme in which an approximate conditional score network is used to generate conditional samples. This approach, known in the literature as *in-painting*, provides high quality results in several *unimodal* application domains [32,53].

By fixing r^{A_1}, r^{A_2}, the KL divergence between $q(z^{A_2} \,|\, \mathcal{C}(r^{A_1}, r^{A_2}), t)$ and $q(z^{A_2} \,|\, r^{A_2}, t)$ quantifies the discrepancy between the true and approximated conditional probabilities. Similarly, the expected KL divergence

$$\Delta = \int q(r,t) \mathrm{KL}[q(z^{A_2} \,|\, \mathcal{C}(r^{A_1}, r^{A_2}), t) \,||\, q(z^{A_2} \,|\, r^{A_2}, t)] \mathrm{d}r \tag{10}$$

provides information about the average discrepancy. Simple manipulations allow this to be recast as a discrepancy in terms of the mutual information $\Delta = I(Z^{A_2}; R_t^{A_1}, R_t^{A_2}) - I(Z^{A_2}; R_t^{A_2})$. Information about Z^{A_2} is contained in $R_t^{A_2}$, as the latter is the result of a diffusion with the former as initial conditions, corresponding to the Markov chain $R_t^{A_2} \to Z^{A_2}$, and in $R_t^{A_1}$ through the Markov chain $Z^{A_2} \to Z^{A_1} \to R_t^{A_1}$. The positive quantity Δ is close to zero whenever the rate of loss of information with respect to the initial conditions is similar for the two subsets A_1 and A_2. In other terms, $\Delta \simeq 0$ whenever the portion $R_t^{A_2}$ of the whole R_t is a sufficient statistic for Z^{A_2}.

The assumptions underlying the approximation are in general not valid in the case of multimodal learning, where the robustness to stochastic perturbations of latent variables corresponding to the various modalities can vary greatly. In Appendix B, our claims

are empirically supported by ample analysis performed on real data showing that our multi-time diffusion approach consistently outperforms in-painting.

5. Experiments

We compared our MLD method to MVAE [21], MMVAE [22], MOPOE [23], Hierarchical Genertive Model (NEXUS) [26], Multi-view Total Correlation Autoencoder (MVTCAE) [27], and MMVAE+ [29], re-implementing all competitors in the same code base as our method and selecting their best hyperparameters as indicated by the authors (see Appendix D for more details). For a fair comparison, we used the same encoder/decoder architecture for all models. For MLD, the score network was implemented using a simple stacked multilayer perceptron (MLP) with skip connections (see Appendix A for more details). MLD was also contrasted with multimodal diffusion-based approaches: [35] in Appendix B and [36] in Section 5.5.

Evaluation metrics: *Coherence* was measured as in [22,23,29], using pretrained classifiers on the generated data and checking the consistency of their outputs. *Generative quality* was computed using the Fréchet Inception Distance (FID) [54] and Fréchet Audio Distance (FAD) [55] scores for images and audio, respectively. Full details on the metrics are included in Appendix C. All results were averaged over five seeds. We report the standard deviations in Appendix E.

Results: Overall, MLD largely outperformed the alternatives from the literature in terms of both coherence and generative quality. The VAE-based models suffered from the coherence–quality tradeoff as well as from modality collapse for highly heterogeneous datasets. We proceed to show this on several standard benchmarks from the multimodal VAE-based literature; see Appendix C for details on the datasets.

5.1. MNIST-SVHN

The first dataset we consider is **MNIST-SVHN** [22], where the two modalities differ in complexity. High variability, noise, and ambiguity make 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). The mixture models, MMVAE and MOPOE, suffer from modality collapse (poor SVHN generation), whereas the product-of-experts models MVAE and MVTCAE generate better-quality samples at the expense of SVHN to MNIST conditional coherence. Joint generation is poor for all VAE models. Interestingly, these models also fail at SVHN self-reconstruction, which we discuss in Appendix E. MLD also achieves the best performance in terms of generation quality, as confirmed by qualitative results (Figure 2) showing, for example, how MLD conditionally generates multiple SVHN digits within one sample given the input MNIST image, whereas the other methods fail to do so.

Table 1. Generation coherence and quality for **MNIST-SVHN** (M: MNIST, S: SVHN). The generation quality is measured in terms of the Fréchet Modality Distance (FMD) for MNIST and FID for SVHN. We report both joint and conditional generation performance results. Bold and underlined numbers indicate the best and second best scores respectively.

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	<u>68.0</u>	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	<u>70.67</u>	13.84	98.13	281.28	53.41
MVTCAE	<u>48.78</u>	<u>81.97</u>	49.78	<u>12.98</u>	**52.92**	69.48	<u>13.55</u>
MMVAE+	17.64	13.23	29.69	26.60	121.77	240.90	35.11
MMVAE+ (K = 10)	41.59	55.3	56.41	19.05	67.13	75.9	18.16
MLD (ours)	**85.22**	**83.79**	**79.13**	**3.93**	<u>56.36</u>	**57.2**	**3.67**

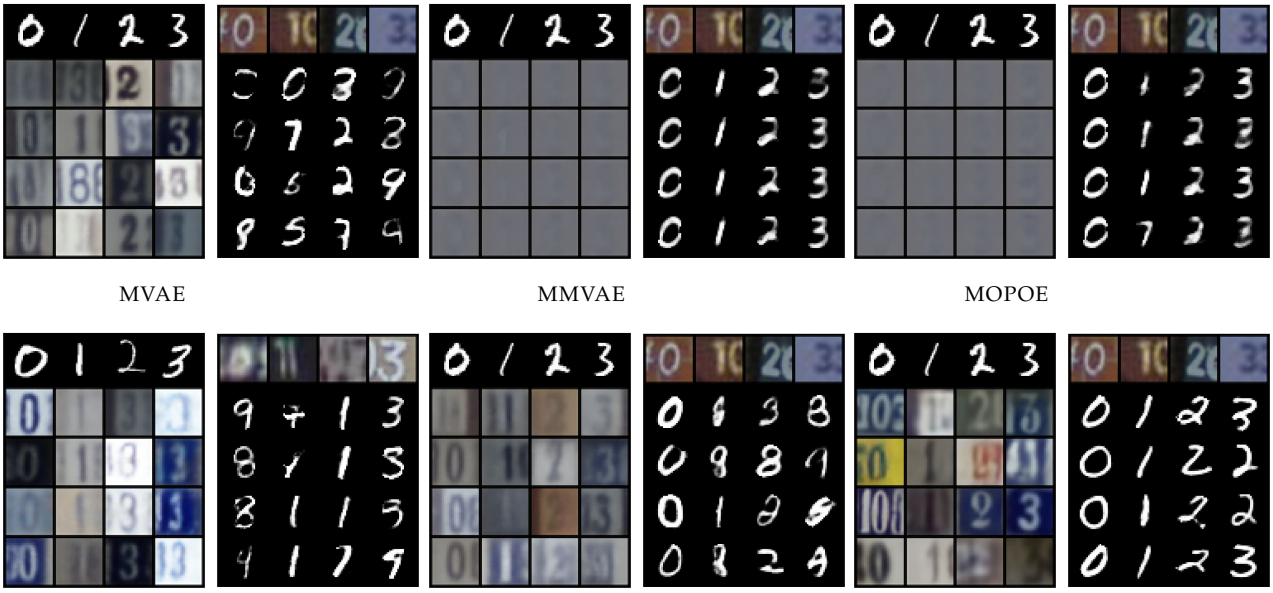


Figure 2. Qualitative results for **MNIST-SVHN**. For each model, we report MNIST to SVHN conditional generation on the left and SVHN to MNIST conditional generation on the right. The conditioning modality is illustrated by the first row, with the generated samples below.

5.2. MHD

The Multimodal Handwritten Digits dataset (**MHD**) [26] contains gray-scale images of digits, the motion trajectory of the handwriting, and the sounds of the spoken digits. In our experiments, we did not use the label as a fourth modality. While the images and trajectories share a good amount of information, the sound modality contains a great deal more modality-specific variation. Consequently, both conditional generation involving the sound modality and joint generation represent challenging tasks. Coherency-wise, (Table 2) MLD outperforms all the competitors, with the biggest difference seen in joint generation and generation from sound to other modalities. On the latter task, MVTCAE performs better than other competitors, but is still worse than MLD. MLD dominates the alternatives in terms of generation quality (Table 3). This is true both for image and sound modalities, for which some VAE-based models struggle to produce high-quality results, demonstrating the limitation of these methods in handling highly heterogeneous modalities. MLD, on the other hand, achieves high generation quality for all modalities, possibly due to the independent training of the autoencoders avoiding interference.

Table 2. Generation coherence (%) for **MHD** (higher is better). Line above refers to the generated modality, while the subset of observed modalities is presented below. Bold and underlined numbers indicate the best and second best scores respectively.

Models	**Joint**	**I (Image)**			**T (Trajectory)**			**S (Sound)**		
		T	**S**	**T,S**	**I**	**S**	**I,S**	**I**	**T**	**I,T**
MVAE	37.77	11.68	26.46	28.4	95.55	26.66	96.58	58.87	10.76	58.16
MMVAE	34.78	**99.7**	69.69	84.74	<u>99.3</u>	85.46	92.39	49.95	50.14	50.17
MOPOE	48.84	<u>99.64</u>	68.67	<u>99.69</u>	99.28	<u>87.42</u>	99.35	50.73	51.5	56.97
NEXUS	26.56	94.58	<u>83.1</u>	95.27	88.51	76.82	93.27	70.06	75.84	89.48
MVTCAE	42.28	99.54	72.05	99.63	99.22	72.03	<u>99.39</u>	<u>92.58</u>	<u>93.07</u>	<u>94.78</u>
MMVAE+	41.67	98.05	84.16	91.88	97.47	81.16	89.31	64.34	65.42	64.88
MMVAE+ (K = 10)	42.60	99.44	**89.75**	94.7	99.44	**89.58**	95.01	87.15	87.99	87.57
MLD (ours)	**98.34**	99.45	<u>88.91</u>	**99.88**	**99.58**	<u>88.92</u>	**99.91**	**97.63**	**97.7**	**98.01**

Table 3. Generation quality for **MHD** in terms of FMD for image and trajectory modalities and FAD for the sound modality (lower is better). Bold and underlined numbers indicate the best and second best scores respectively.

Models	I (Image)				T (Trajectory)				S (Sound)			
	Joint	T	S	T,S	Joint	I	S	I,S	Joint	I	T	I,T
MVAE	<u>94.9</u>	93.73	92.55	91.08	39.51	20.42	38.77	19.25	14.14	<u>14.13</u>	14.08	14.17
MMVAE	224.01	22.6	789.12	170.41	16.52	**0.5**	30.39	6.07	22.8	22.61	23.72	23.01
MOPOE	147.81	16.29	838.38	15.89	<u>13.92</u>	<u>0.52</u>	33.38	**0.53**	18.53	24.11	24.1	23.93
NEXUS	281.76	116.65	282.34	117.24	18.59	6.67	33.01	7.54	<u>13.99</u>	19.52	18.71	16.3
MVTCAE	121.85	<u>5.34</u>	<u>54.57</u>	<u>3.16</u>	19.49	0.62	<u>13.65</u>	0.75	15.88	14.22	<u>14.02</u>	<u>13.96</u>
MMVAE+	97.19	2.80	128.56	114.3	22.37	1.21	21.74	15.2	16.12	17.31	17.92	17.56
MMVAE+ (K = 10)	85.98	1.83	70.72	62.43	21.10	1.38	8.52	7.22	14.58	14.33	14.34	14.32
MLD	**7.98**	**1.7**	**4.54**	**1.84**	**3.18**	0.83	**2.07**	<u>0.6</u>	**2.39**	**2.31**	**2.33**	**2.29**

5.3. POLYMNIST

The **POLYMNIST** dataset [23] consists of five modalities synthetically generated using MNIST digits and varying the background images. The homogeneous nature of the modalities is expected to mitigate gradient conflict issues in VAE-based models and consequently reduce modality collapse. However, MLD still outperforms all alternatives, as shown in Figures 3 and 4. Concerning generation coherence, MLD achieves the best performance in all cases, with the one exception of a single observed modality. On the qualitative performance side, not only is MLD superior to all alternatives, its results are stable when more modalities are considered, a capability that not all competitors share.

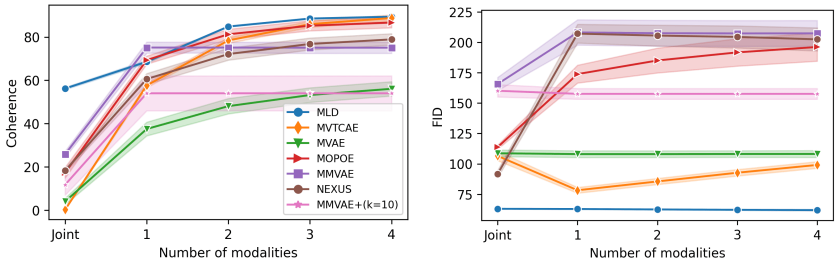


Figure 3. Performance results for **POLYMNIST** as a function of the number of inputs. (**Right**): Generative coherence (% ↑). (**Left**): Generative quality in terms of FID (↓). We report the average performance following the leave-one-out strategy (see Appendix C).

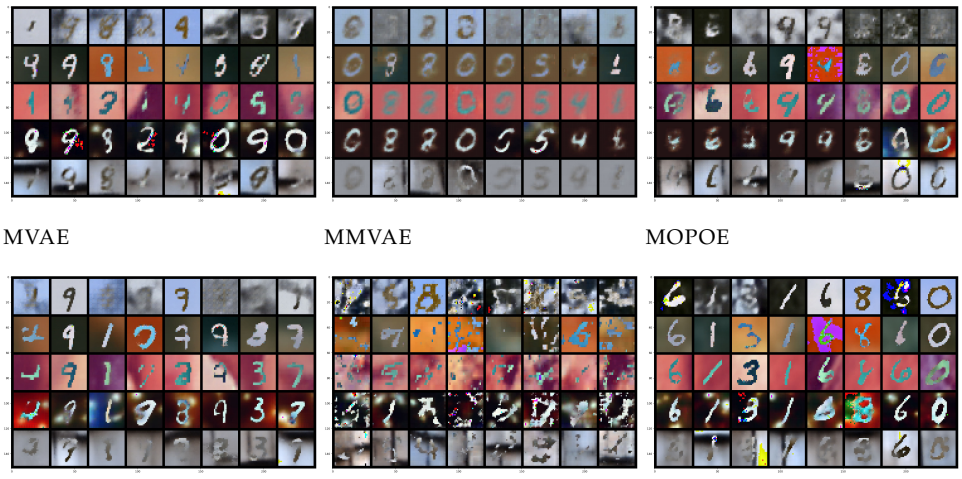


Figure 4. Joint generation qualitative results for **POLYMNIST** across the five modalities.

5.4. CUB

Next, we explored the Caltech Birds **CUB** [22] dataset, following the experimental protocol in [24] using real bird images instead of ResNet-features as in [22]. Figure 5 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 considerably improves the generative performance of MLD, suggesting that careful engineering applied to modality-specific autoencoders is a promising avenue for future work. We report quantitative results in Appendix E, where we show the generation quality FID metric. Due to the unavailability of the labels in this dataset, the coherence evaluation performed with the previous datasets was not possible. Thus, we resorted to CLIP-Score (CLIP-S) [56], an image-captioning metric. Despite its limitations for the considered dataset [57], CLIP-S shows that MLD outperforms all competitors.

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

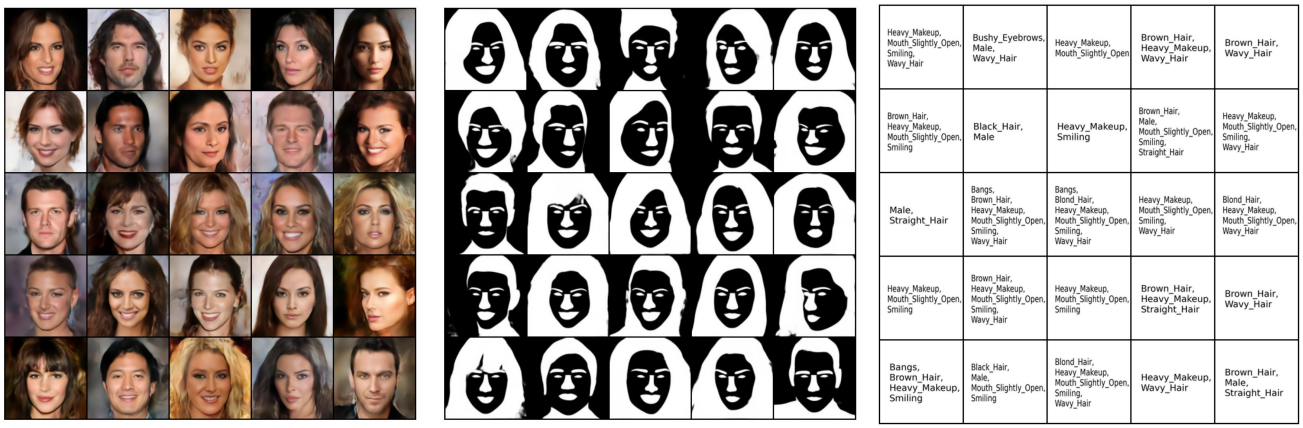
Figure 5. Qualitative results on the **CUB** dataset, with the caption used as the condition to generate the bird images. **MLD*** denotes the version of our method using a powerful image autoencoder.

5.5. CelebAMask-HQ

Finally, we considered the CelebAMask-HQ dataset [58], which consists of three modalities: face images, each having a segmentation mask and text attributes. We followed the same experimental protocol as in [36], including the autoencoder base architecture. The image generation quality was evaluated in terms of FID score. The attributes and the mask, both having binary values, were evaluated against the ground truth in terms of the *F*1 score. The competitors' performance results are reported from [36]. The quantitative results in Table 4 show that MLD outperforms the competitors in terms of generation quality. Our method achieves the best *F*1 score in generation of the attribute modalities given the image and mask modalities. In mask generation, MOPOE and MVTCAE achieve the best performance, with MLD achieving the second-best performance in mask generation conditioning on both the image and attribute modalities. Overall, MLD stands out with the best image quality generation, while being on par with the competition in terms of mask and attribute generation coherence. Figure 6 shows the qualitative results for MLD on the joint generation task. It can be observed that our method succeeds at generating all three modalities with high coherence and quality. The same observation is valid for the conditional generation tasks (see Figures 7–9).

Table 4. Quantitative results on the CelebAMask-HQ dataset. Performance is measured in terms of the FID (↓) and F1 score (↑). The first row shows the generated modality, while the second row shows the modalities used as conditions. Supervised classifier designates a classifier performance to predict the attributes or the mask from an image. Bold numbers indicate the best scores.

Models	Attributes		Image				Mask	
	Img + Mask F1	Img F1	Att + Mask FID	Mask FID	Att FID	Joint FID	Img + Att F1	Img F1
SBM-RAE [36]	0.62	0.6	84.9	86.4	85.6	84.2	0.83	0.82
SBM-RAE-C [36]	0.66	0.64	83.6	82.8	83.1	84.2	0.83	0.82
SBM-VAE [36]	0.62	0.58	81.6	81.9	78.7	79.1	0.83	0.83
SBM-VAE-C [36]	0.69	0.66	82.4	81.7	76.3	79.1	0.84	0.84
MOPOE	0.68	**0.71**	114.9	101.1	186.8	164.8	0.85	**0.92**
MVTCAE	0.71	0.69	94	84.2	87.2	162.2	**0.89**	0.89
MMVAE+	0.64	0.61	133	97.3	153	103.7	0.82	0.89
Supervised classifier		0.79					0.94	
MLD (ours)	**0.72**	0.69	**52.75**	**51.73**	**53.09**	**54.27**	0.87	0.87



(**a**) Image (**b**) Mask (**c**) Attributes

Figure 6. Joint (unconditional) generation: qualitative results of **MLD** on CelebAMask-HQ.

Figure 7. (Attributes → Image). Conditional generation of **MLD** on CelebAMask-HQ. The first column on the left presents the conditioning modalities, while several conditionally generated samples are displayed on the right.

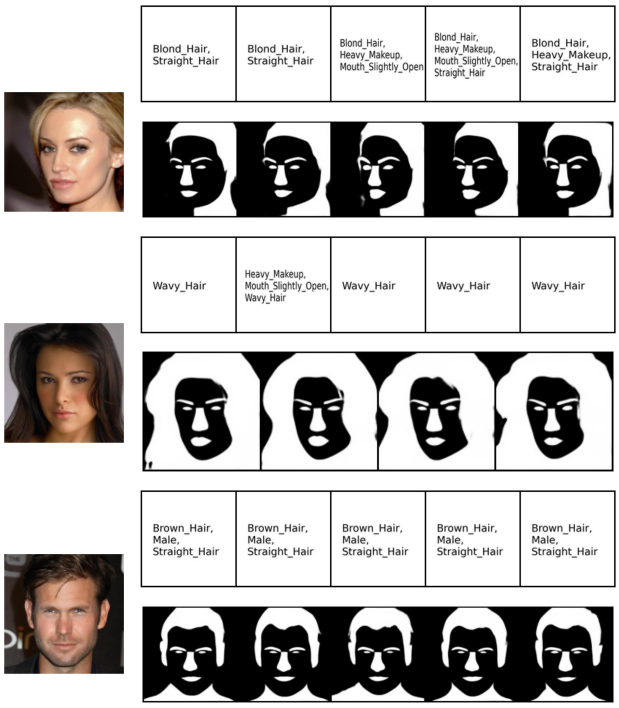


Figure 8. (Image → Attribute, Mask). Conditional generation of **MLD** on CelebAMask-HQ. The first column on the left presents the conditioning modalities, while several conditionally generated samples are displayed on the right.

Figure 9. (Attributes, Mask → Image). Conditional generation of **MLD** on CelebAMask-HQ. The two columns on the left present the conditioning modalities, while several conditionally generated samples are displayed on the right.

6. Conclusions and Limitations

We have presented a new multimodal generative model, Multimodal Latent Diffusion (MLD), to address the well known coherence–quality tradeoff that is inherent in existing multimodal VAE-based models. MLD uses a set of independently trained unimodal deterministic autoencoders. The generative properties of our model stem from a masked diffusion process that operates on latent variables. In addition, we have developed a new multi-time training method to learn the conditional score network for multimodal diffusion. An extensive experimental campaign on various real-life datasets provides compelling evidence of the effectiveness of MLD for multimodal generative modeling. In all scenarios, including cases with loosely correlated modalities and high-resolution datasets, MLD consistently outperforms state-of-the-art alternatives. A limitation of our approach stems from the simple nature of encoder/decoder architectures. Focusing on more specialized, complex, and tailor-made encoder/decoder architectures might be necessary when moving

to higher-resolution data. As for all generative models, ours could be misused to produce misinformation. We believe, however, that the benefits of multimodal generative models outweigh their potential misuses.

Author Contributions: Conceptualization, M.B., G.F. and P.M.; Methodology, M.B., G.F. and P.M.; Software, M.B.; Validation, M.B., G.F. and P.M.; Investigation, M.B., G.F. and P.M.; Writing—original draft, M.B., G.F. and P.M.; Writing—review & editing, M.B., G.F. and P.M.; Supervision, G.F. and P.M. All authors have read and agreed to the published version of the manuscript.

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

Data Availability Statement: All used datasets are publicly available. Our code is available at https://github.com/MustaphaBounoua/MLD.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Diffusion in the Multimodal Latent Space

In this appendix, we provide additional technical details of MLD. We first discuss a naive approach based on *in-painting* which uses only the unconditional score network for both joint and conditional generation. We also discuss an alternative training scheme based on a work from the caption-text translation literature [35]. Finally, we provide extra technical details for the score network architecture and sampling technique.

Appendix A.1. Modality Auto-Encoders

Each of the 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 the same size to allow for the definition of a joint latent space.

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

Appendix A.2. Multimodal Diffusion SDE

In Section 3, we presented our multimodal latent diffusion process allowing multimodal 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) [32]. 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 [32,59], we set $\beta_{min} = 0.1$ and $\beta_{max} = 20$. With this configuration, and by substitution of Equation (4), we obtain the following forward SDE:

$$\mathrm{d}R_t = -\frac{1}{2}\beta(t)R_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}W_t, \qquad t \in [0, T]. \tag{A1}$$

The corresponding perturbation kernel is provided by

$$q(r|z,t) = \mathcal{N}(r; e^{-\frac{1}{4}t^2(\beta_{max}-\beta_{min})-\frac{1}{2}t\beta_{min}}z, (1 - e^{-\frac{1}{2}t^2(\beta_{max}-\beta_{min})-t\beta_{min}})\mathbf{I}). \tag{A2}$$

The marginal score $\nabla \log q(R_t, t)$ is approximated by a score network $s_\chi(R_t, t)$, the parameters χ of which can be optimized by minimizing the ELBO in Equation (6), where we found that using the same re-scaling as in [32] is more stable.

The reverse process is described by a different SDE (Equation (5)). When using a variance-preserving SDE, Equation (5) specializes in

$$\mathrm{d}R_t = \left[\frac{1}{2}\beta(T-t)R_t + \beta(T-t)\nabla \log q(R_t, T-t)\right]\mathrm{d}t + \sqrt{\beta(T-t)}\mathrm{d}W_t, \tag{A3}$$

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

When the parametric score network has been optimized through the simulation of Equation (A3), sampling $R_T \sim q_\psi(r)$ becomes possible, allowing **joint generation**. A numerical SDE solver can be used to sample R_T, which can then be fed to the modality-specific decoders to jointly sample a set of $\hat{X} = \{d^i_\theta(R^i_T)\}^M_{i=0}$. As explained in Section 4.3, the use of the unconditional score network $s_\chi(R_t, t)$ allows for **conditional generation** through the approximation described in [32].

As described in Algorithm A1, we can generate a set of modalities A_1 conditioned on the available set of modalities A_2. 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), and the reverse diffusion SDE in Equation (A3) is discretized using a finite time step $\Delta t = {}^T/_N$, starting from $t = 0$ and iterating until $t \approx T$. At each iteration, the available portion of the latent space is diffused and brought to the same noise level as $R_t^{A_1}$, allowing for the use of the unconditional score network. Lastly, the reverse diffusion update is performed. This process is repeated until arriving at $t \approx T$ and obtaining $R_T^{A_1} = \hat{Z}^{A_1}$, which can be decoded to recover $\hat{x}^{A_1}$. Note that this joint generation can be seen as a special case of Algorithm A1 with $A_2 = \emptyset$. We name this first approach Multi-modal Latent Diffusion with In-painting (MLD IN-PAINT), and provide extensive comparison with our MLD method in Appendix B.

Algorithm A1: MLD IN-PAINT conditional generation

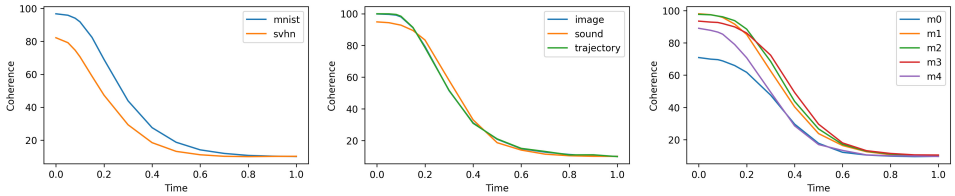
Data: $x^{A_2} = \{x^i\}_{i \in A_2}$
$z^{A_2} \leftarrow \{e_{\phi_i}(x^i)\}_{i \in A_2}$ `// Encode the available modalities` X `into their latent space`
$A_1 \leftarrow \{1, \ldots, M\} \setminus A_2$ `// The set of modalities to generate`
$R_0 \leftarrow \mathcal{C}(R_0^{A_1}, z^{A_2}), \quad R_0^{A_1} \sim \rho(r^{A_1})$ `// Compose the initial state`
$R \leftarrow R_0$
$\Delta t \leftarrow {}^T/_N$
for $n = 0$ ***to*** $N - 1$ **do**
  $t' \leftarrow T - n\,\Delta t$
  $\bar{R} \sim q(r|R_0, t')$ `// Diffuse the available portion of the latent space(Equation (A2))`
  $R \leftarrow m(A_1) \odot R + (1 - m(A_1)) \odot \bar{R}$
  $\epsilon \sim \mathcal{N}(0; I)$ **if** $n < (N-1)$ **else** $\epsilon = 0$
  $\Delta R \leftarrow \Delta t\left[\frac{1}{2}\beta(t')R + \beta(t')s_\chi(R, t')\right] + \sqrt{\beta(t')\Delta t}\epsilon$
  $R \leftarrow R + \Delta R$ `// The Euler-Maruyama update step`
end
$\hat{z}^{A_1} \leftarrow R^{A_1}$
Return $\hat{X}^{A_1} = \{d^i_\theta(\hat{z}^i)\}_{i \in A_1}$

As discussed in Section 4.3, the approximation enabling the in-painting approach can be efficient in several domains; however, its generalization to the multimodal latent space scenario is not trivial. We argue that this is due to the heterogeneity of modalities. which induce different characteristics on the part of the latent spaces. For different modality-specific latent spaces, the loss of information ratio can vary through the diffusion process. We verify this hypothesis by the following experiment.

Latent space robustness against diffusion perturbation.

We analyse the effect of the forward diffusion perturbation on the latent space through time. We encode the modalities using their respective encoders to obtain their latent space $Z = [e_{\psi^1}(X^1) \dots e_{\psi^M}(X^M)]$. Given a time $t \in [0, T]$, we diffuse the different latent spaces by applying Equation (A2) to obtain $R_t \sim q(r|z, t)$, with R_t being the perturbed version of the latent space at time t. We feed the modality-specific decoders with the perturbed latent space $\hat{X}_t = \{d^i_\theta(R^i_t)\}^M_{i=1}$, with $\hat{X}_t$ being the output modalities generated using the perturbed latent space. To evaluate the information loss induced by the diffusion process on the different modalities, we assess the coherence preservation in the reconstructed modalities $\hat{X}_t$ by computing the coherence (in %) as done in Section 5.

We expect to obtain high coherence results for $t \approx 0$ when compared to $t \approx T$, asthe information in the latent space is more preserved at the beginning of the diffusion process than at the last phase of the forward SDE, where all dependencies on initial conditions vanish. Figure A1 shows the coherence as a function of the diffusion time $t \in [0, 1]$ for different modalities across multiple datasets. It can be observed that, within the same dataset, certain modalities stand out with a specific level of robustness (using the coherence level as a proxy) against the diffusion perturbation in comparison with the remaining modalities from the same dataset. For instance, we remark that SVHN is less robust than MNIST, which should manifest in underperformance of SVHN-to-MNIST conditional generation. We verify this intuition in Appendix B.



(**a**) MNIST-SVHN (**b**) MHD (**c**) POLYMNIST

Figure A1. Coherence as a function of the diffusion process time for three datasets. Diffusion perturbation is applied on the modalities' latent space after element-wise normalization.

Appendix A.3. Multi-Time Masked Multimodal SDE

in Section 4, we proposed a multi-time masked diffusion process to learn a score network capable of both conditional and joint generation.

Algorithm A2 presents the pseudo-code for the multi-time masked training. The masked diffusion process is applied following randomization with probability d. First, a subset of modalities A_2 is selected randomly to be the conditioning modalities, with A_1 the remaining set of modalities to make up 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 A2, the portion of the latent space corresponding to the subset A_2 is not diffused and is forced to be equal to $R^{A_2}_0 = z^{A_2}$. 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 $(1 - d)$, all the modalities are diffused at the same time and $A_2 = \emptyset$. 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 frozen portions of the latent space:

$$\Omega(A_1, A_2) = 1 + \frac{dim(A_2)}{dim(A_1)}, \tag{A4}$$

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

Algorithm A2: 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`
$A_2 \sim \nu$ `//` ν `depends on the parameter` d
$A_1 \leftarrow \{1, \ldots, M\} \setminus A_2$
$t \sim \mathcal{U}[0, T]$
$R \sim q(r|Z, t)$ `// Diffuse the available portion of the latent space(Equation (A2))`
$R \leftarrow m(A_1) \odot R + (1 - m(A_1)) \odot Z$ `// Masked diffusion`
$\tau(A_1, t) \leftarrow [\mathbb{1}(1 \in A_1)t, \ldots, \mathbb{1}(M \in A_1)t]$ `// Construct the multi time vector`
Return $\nabla_\chi \Big\{ \Omega(A_1, A_2) \quad \big\| m(A_1) \odot \quad \big[s_\chi(R, \tau(A_1, t)) - \nabla \log q(R, t|z^{A_2}) \big] \big\|_2^2 \quad \Big\}$

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

$$s_\chi(R_t, \tau(A_1, t)) \sim \nabla \log q(R_t, t \,|\, z^{A_2})). \tag{A5}$$

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

$$s_\chi(R_t, \tau(A_1, t)) \sim \nabla \log q(R_t, t) \quad , A_1 = \{1, \ldots, M\}. \tag{A6}$$

Algorithm A3 describes the reverse conditional generation pseudo-code. It is pertinent to compare this algorithm with Algorithm A1. The main difference resides in the use of the multi-time score network to enable conditional generation, with the multi-time vector playing the role of the time information and conditioning signal. On the other hand, in Algorithm A1, we do not have a conditional score network; therefore, we resort to the approximation from Section 4.3 and use the unconditional score.

Algorithm A3: 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`
$A_1 \leftarrow \{1, \ldots, M\} \setminus A_2$ `// The set of modalities to be generated`
$R_0 \leftarrow \mathcal{C}(R_0^{A_1}, z^{A_2}), \quad R_0^{A_1} \sim \rho(r^{A_1})$ `// Compose the initial latent space`
$R \leftarrow R_0$
$\Delta t \leftarrow T/N$
for $n = 0$ ***to*** $N - 1$ **do**
  $t' \leftarrow T - n\,\Delta t$
  $\tau(A_1, t') \leftarrow [\mathbb{1}(1 \in A_1)t', \ldots, \mathbb{1}(M \in A_1)t']$ `// Construct the multi-time vector`
  $\epsilon \sim \mathcal{N}(0; I)$ **if** $n < N$ **else** $\epsilon = 0$
  $\Delta R \leftarrow \Delta t \Big[\frac{1}{2}\beta(t')R + \beta(t')s_\chi(R, \tau(A_1, t')) \Big] + \sqrt{\beta(t')\Delta t}\epsilon$
  $R \leftarrow R + \Delta R$ `// 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 $\hat{X}^{A_1} = \{d_\theta^i(\hat{z}^i)\}_{i \in A_1}$

Appendix A.4. Uni-Diffuser Training

The work presented in [35] is specialized for an image–caption application. The approach is based on a multimodal diffusion model applied to a unified latent embedding obtained via pretrained autoencoders and incorporates pretrained models (CLIP [60] and GPT-2 [61]). The unified latent space is composed of an image embedding, a CLIP image embedding, and a CLIP text embedding. Note that the CLIP model is pretrained on (image–text) pairs of multimodal data, which is expected to enhance the generative performance. Because it is non-trivial to have a jointly trained encoder similar to CLIP for any type of modality, the evaluation of this model on different modalities across different datasets (e.g., including audio) is not an easy task.

To compare to this work, we adapted the training scheme presented in [35] to our MLD method. Instead of applying a masked multimodal SDE to train the score network, every portion of the latent space was diffused according to a different time $t^i \sim \mathcal{U}(0,1)$; therefore, the multi-time vector fed to the score network was $\tau(t) = [t^0 \sim \mathcal{U}(0,1), \ldots, t^M \sim \mathcal{U}(0,1)]$. For fairness, we used the same score network and reverse process sampler as was used for our MLD version with multi-time training; we call this variant Multi-modal Latent Diffusion UniDiffuser (MLD UNI).

Appendix A.5. Technical Details

Appendix A.5.1. Sampling Schedule

We used the sampling schedule proposed in [53], which has been shown to improve the coherence of conditional and joint generation. We used the best parameters suggested by the authors: $N = 250$ time steps applied $r = 10$ resampling times with jump size $j = 10$. For readability, in Algorithms A1 and A3 we present pseudo-code with a linear sampling schedule which can be easily adapted to any other schedule.

Appendix A.5.2. Training the Score Network

Inspired by the architecture from [62], we use simple Residual MLP blocks with skip connections as our score network (see Figure A2). We fix the **width** and **number of blocks** proportionally to the number of the modalities and the latent space size. As in [63], we use the Exponential moving average (EMA) of the model parameters with a momentum parameter $m = 0.999$.

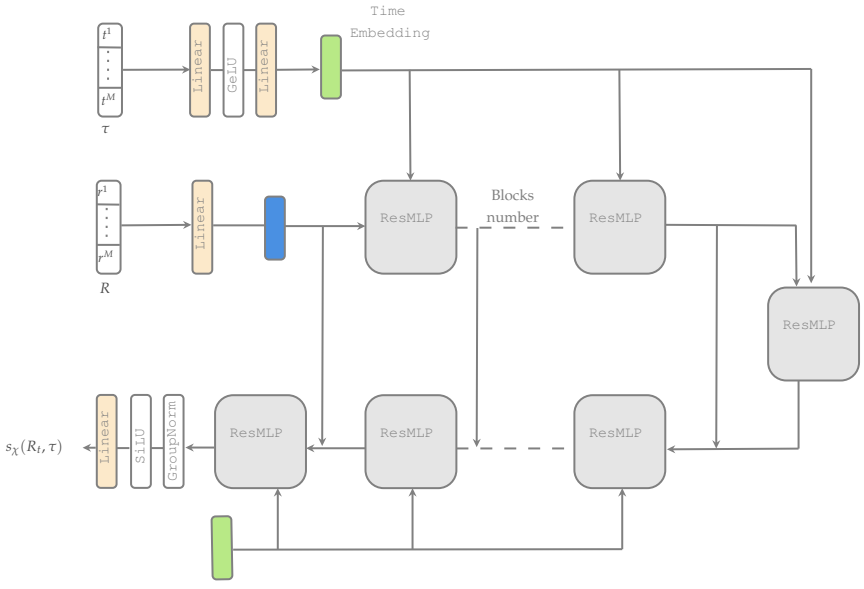


Figure A2. Score network s_χ architecture used in our MLD implementation. The residual MLP block architecture is shown in Figure A3.

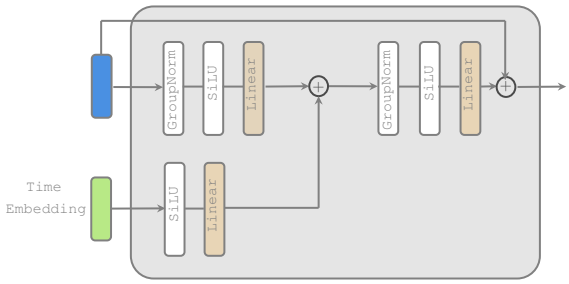


Figure A3. Architecture of the ResMLP block.

Appendix B. MLD Ablation Study

In this section, we compare MLD with two variants presented in Appendix A: MLD IN-PAINT, a naive approach without our proposed *multi-time masked* SDE, and MLD UNI, a variant of our method using the same training scheme from [35]. In addition, we analyze the effect of the randomization parameter d on the performance of MLD through an ablation study.

Appendix B.1. MLD and Its Variants

Table A1 summarizes the different approaches adopted in each variant. All the considered models share the same deterministic autoencoders trained during the first stage.

For fairness, our evaluation was carried out using the same configuration and code basis as MLD. This included the autoencoder architectures and latent space size (similar to Section 5). The same score network (Figure A2) was used across experiments, with MLD IN-PAINT using the same architecture with one time dimension instead of the multi-time vector. In all the variants, joint and conditional generation were conducted using the same reverse sampling schedule described in Appendix A.

Table A1. Ablation study of MLD and its variants.

Model	Multi-Time Diffusion	Training	Conditional and Joint Generation
MLD IN-PAINT	×	Equation (6)	Algorithm A1
MLD UNI	✓	[35]	Algorithm A3
MLD	✓	Algorithm A2	Algorithm A3

Appendix B.1.1. Results

In certain cases, the MLD variants were able to match the joint generation performance of MLD; however, overall they were less efficient and had noticeable weaknesses. MLD IN-PAINT underperforms on conditional generation when considering relatively complex modalities, while MLD UNI is not able to leverage the presence of multiple modalities to improve cross-generation, especially for datasets with a large number of modalities. On the other hand, MLD is able to overcome these limitations.

Appendix B.1.2. MNIST-SVHN

In Table A2, MLD achieves the best results and dominates cross=generation performance. It can be observed that MLD IN-PAINT lacks coherence for SVHN-to-MNIST conditional generation, a result we expected based on our analysis of the experiment in Figure A1. MLD UNI, despite the use of a multi-time diffusion process, underperforms our method, which indicates the effectiveness of our masked diffusion process in learning the conditional score network. Because all of the models used the same deterministic autoencoders, their observed generative quality performances are relatively similar (see Figure A4 for qualitative results).

Table A2. Generation coherence and quality for MNIST-SVHN (M stands for MNIST and S for SVHN). The generation quality is measured in terms of FMD for MNIST and FID for SVHN. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Coherence (%↑)			Quality (↓)			
	Joint	M → S	S → M	Joint (M)	Joint (S)	M → S	S → M
MLD-Inpaint	**85.53** ± 0.22	<u>81.76</u> ± 0.23	63.28 ± 1.16	**3.85** ± 0.02	60.86 ± 1.27	59.86 ± 1.18	**3.55** ± 0.11
MLD-Uni	82.19 ± 0.97	79.31 ± 1.21	<u>72.78</u> ± 1.81	4.1 ± 0.17	57.41 ± 1.43	<u>57.84</u> ± 1.57	4.84 ± 0.28
MLD	<u>85.22</u> ± 0.5	**83.79** ± 0.62	**79.13** ± 0.38	<u>3.93</u> ± 0.12	**56.36** ± 1.63	**57.2** ± 1.47	<u>3.67</u> ± 0.14

MLD IN-PAINT MLD UNI **MLD**

Figure A4. Qualitative results for **MNIST-SVHN**. For each model, we report MNIST-to-SVHN conditional generation on the left and SVHN-to-MNIST conditional generation on the right.

Appendix B.1.3. MHD

Table A3 shows the performance results for the MHD dataset in terms of generative coherence. MLD achieves the best joint generation coherence, and, dominates the cross-generation coherence results along with MLD UNI. MLD IN-PAINT shows a lack of coherence when conditioning on the sound modality alone, which is a predictable result, as this is a more difficult configuration because the sound modality is loosely correlated to other modalities. It can be observed that MLD IN-PAINT performs worse than the two other alternatives when conditioned on the trajectory modality, which is the smallest modality in terms of latent size. This indicates another limitation of the naive approach regarding coherent generation when handling different latent spaces sizes, a weakness that our MLD method overcomes. Table A4 presents the qualitative generative performance results, which are homogeneous across the variants, with MLD achieving either the best or second-best performance.

Table A3. Generation coherence (%↑) for MHD (higher is better). The line above refers to the generated modality, while the subset of observed modalities is presented below. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Joint	I (Image)			T (Trajectory)				S (Sound)		
		T	S	T,S	I	S	I,S	I	T	I,T
MLD-Inpaint	96.88 ± 0.35	63.9 ± 1.7	56.52 ± 1.89	95.83 ± 0.48	<u>99.58</u> ± 0.1	56.51 ± 1.89	<u>99.89</u> ± 0.04	95.81 ± 0.25	56.51 ± 1.89	96.38 ± 0.35
MLD-Uni	<u>97.69</u> ± 0.26	**99.91** ± 0.04	**89.87** ± 0.38	**99.92** ± 0.04	**99.68** ± 0.1	**89.78** ± 0.45	99.38 ± 0.31	<u>97.54</u> ± 0.2	<u>97.65</u> ± 0.41	<u>97.79</u> ± 0.41
MLD	**98.34** ± 0.22	99.45 ± 0.09	<u>88.91</u> ± 0.54	<u>99.88</u> ± 0.04	<u>99.58</u> ± 0.03	<u>88.92</u> ± 0.53	**99.91** ± 0.02	**97.63** ± 0.14	**97.7** ± 0.34	**98.01** ± 0.21

Table A4. Generation quality for MHD. The metrics reported are FMD for the image and trajectory modalities and FAD for the sound modality (lower is better). Bold and underlined numbers indicate the best and second best scores respectively.

Models	I (Image)				T (Trajectory)				S (Sound)			
	Joint	T	S	T,S	Joint	I	S	I,S	Joint	I	T	I,T
MLD-Inpaint	5.35 ± 1.35	6.23 ± 1.13	<u>4.76</u> ± 0.68	3.53 ± 0.36	**1.59** ± 0.12	**0.6** ± 0.05	**1.81** ± 0.13	**0.54** ± 0.06	2.41 ± 0.07	2.5 ± 0.04	2.52 ± 0.02	2.49 ± 0.05
MLD-Uni	**7.91** ± 2.2	**1.65** ± 0.33	6.29 ± 1.38	<u>3.06</u> ± 0.54	<u>2.53</u> ± 0.5	1.18 ± 0.26	3.18 ± 0.77	2.84 ± 1.14	**2.11** ± 0.08	**2.25** ± 0.05	**2.1** ± 0.0	**2.15** ± 0.01
MLD	<u>7.98</u> ± 1.41	<u>1.7</u> ± 0.14	**4.54** ± 0.45	**1.84** ± 0.27	3.18 ± 0.18	<u>0.83</u> ± 0.03	<u>2.07</u> ± 0.26	<u>0.6</u> ± 0.05	<u>2.39</u> ± 0.1	<u>2.31</u> ± 0.07	<u>2.33</u> ± 0.11	<u>2.29</u> ± 0.06

Appendix B.1.4. POLYMNIST

In Figure A5, we note the superiority of MLD in both generative coherence and quality. MLD-Uni is not able to leverage the presence of a large number of modalities in conditional generation coherence. Interestingly, an increase in the number of input modalities negatively impacts the performance of MLD UNI.

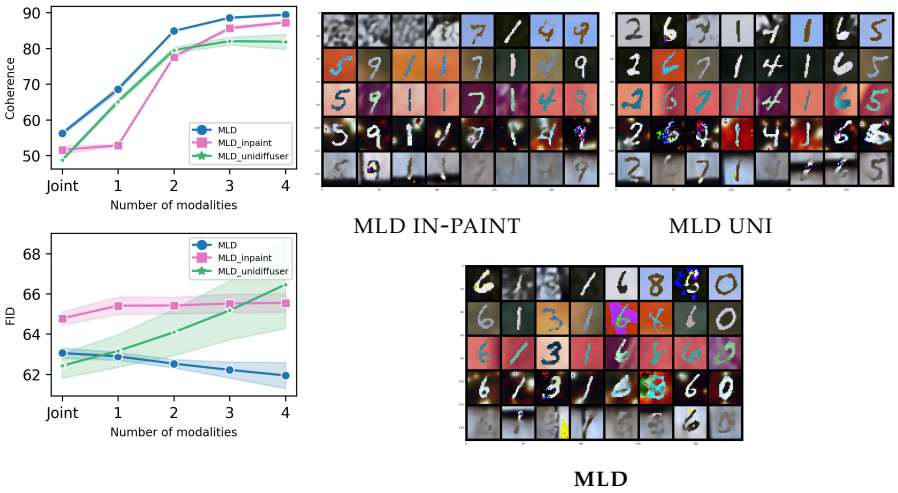


Figure A5. Results for the **POLYMNIST** dataset. (**Left**): a comparison of the generative coherence (% ↑) and quality in terms of FID (↓)) as a function of the number of modality inputs. We report the average performance following the leave-one-out strategy (see Appendix C). (**Right**): qualitative results for joint generation of the five modalities.

Appendix B.1.5. CUB

Figure A6 shows the qualitative results for caption-to-image conditional generation. All of the variants are based on the same first-stage autoencoders, and the generative performance is comparable in terms of quality.

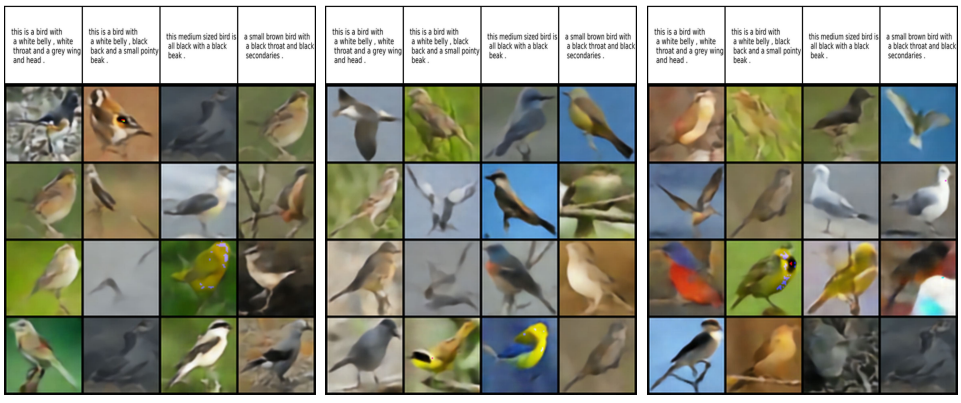


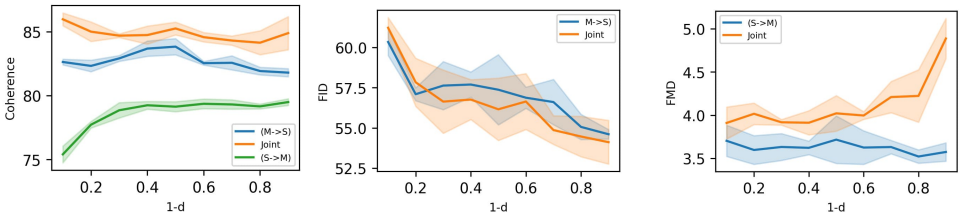
Figure A6. Qualitative results on the **CUB** dataset. Captions were used as the condition to generate the bird images.

Appendix B.2. Randomization d-Ablation Study

The d parameter controls the randomization of the *multi-time masked diffusion process* during training in Algorithm A2. With probability d, the concatenated latent space corresponding to all the modalities is diffused at the same time. With probability $(1 - d)$, a portion of the latent space corresponding to a random subset of the modalities is not diffused and is frozen during the training step. To study the d parameter and its effect on the performance of our MLD model, we used $d \in \{0.1, \ldots, 0.9\}$. Figure A7 shows the

results of the d-ablation study on the **MNIST-SVHN** dataset. We report the performance results averaged over five independent seeds as a function of the probability $(1 - d)$: **Left** shows the conditional and joint coherence for the **MNIST-SVHN** dataset; **Middle** shows the quality performance in terms of FID for SVHN generation; and **Right** shows the quality performance in terms of FMD for MNIST generation.

It can be observed that higher values for $1 - d$, indicating a greater probability of applying *multi-time masked diffusion*, improve the coherence of SVHN-to-MNIST conditional generation. This confirms that masked multi-time training enables better conditional generation. Overall, on the **MNIST-SVHN** dataset, MLD shows weak sensibility to the d parameter whenever the value of $d \in [0.2, 0.7]$.



(**a**) MNIST-SVHN: Coherence (%↑) (**b**) SVHN: FID (↓) (**c**) MNIST: FMD(↓)

Figure A7. Results of the ablation study for the randomization parameter d on the **MNIST-SVHN** dataset.

Appendix C. Datasets and Evaluation Protocol

Appendix C.1. Dataset Description

MNIST-SVHN [22] is constructed using pairs of MNIST and SVHN sharing the same digit class (see Figure A8a). Each instance of a digit class (in either dataset) is randomly paired with 20 instances of the same digit class from the other dataset. SVHN modality samples are obtained from house numbers in Google Street View images, and are 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 dataset for multimodal generative models is to learn to extract digit number and reconstruct a coherent MNIST modality.

MHD [26] is composed of three modalities: synthetically generated images and motion trajectories of handwritten digits associated with their speech sounds. The images are gray-scale $1 \times 28 \times 28$, and the handwriting trajectories are represented by a 1×200 vector. The spoken digits sounds are $1s$ audio clips processed as Mel-Spectrograms, and are constructed with a hopping window of 512 ms with 128 Mel Bins, resulting in a $1 \times 128 \times 32$ representation. This benchmark is the closest to a real-world scenario involving multimodal sensors because of the presence of three completely different modalities, with the audio modality representing a complex data type. Therefore, similar to SVHN, the conditional generation of sound to coherent images or trajectories represents a challenging use case.

POLYMNIST [23] is a version of the MNIST dataset extended to five modalities. Each modality is constructed using a random set of MNIST digits with an overlay over a random crop from a modality-specific three-channel image background. This synthetic generated dataset allows for the evaluating the scalability of multimodal generative models to large number of modalities. Although this dataset is composed only of images, the different textures of different modality-specific backgrounds results in differing levels of difficulty. In Figure A8c, the digits are more difficult to distinguish in modalities 1 and 5 than in the other modalities.

CUB [22] is comprised of bird images and associated text captions. In [22], a simplified version based on precomputed ResNet-features was used. Following [24], we conducted all of our experiments on the real image data instead. Each image from the 11,788 photos

of birds from Caltech-Birds [64] was resized to a $3 \times 64 \times 64$ image and coupled with ten textual descriptions of the respective bird (see Figure A8d).

CelebAHQ-mask consists of three modalities: face images, each with a segmentation mask and attributes. We took into account 18 out of 40 attributes from the original dataset and resized the images to 128×128 resolution, as was done in [21,36].

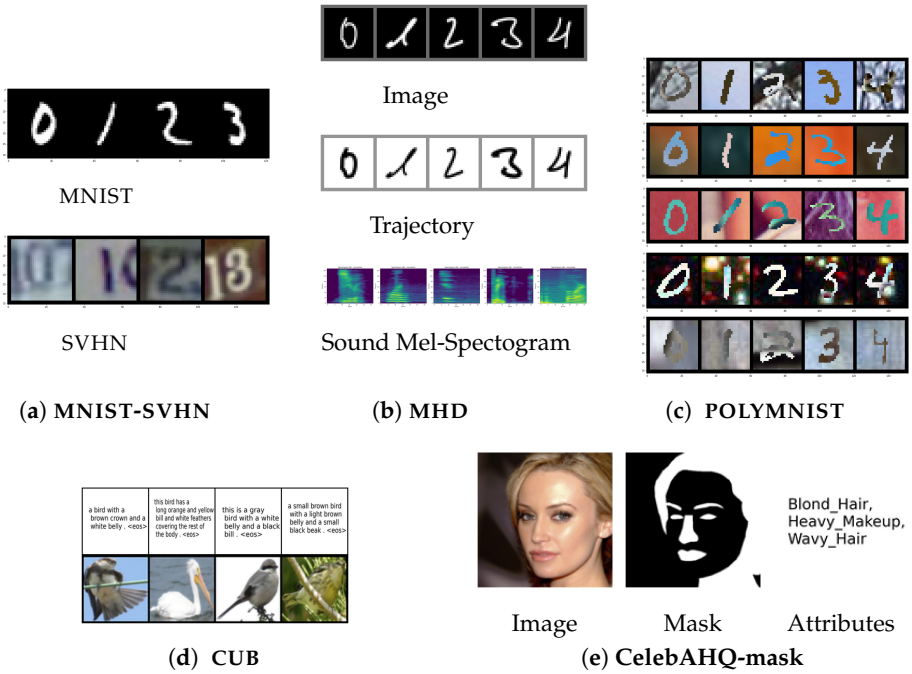


Figure A8. Illustrative example of the datasets used for evaluation.

Appendix C.2. Evaluation Metrics

The multimodal generative models were evaluated in terms of their generative coherence and quality.

Appendix C.2.1. Generation Coherence

We measured *coherence* by verifying that generated data for both joint and conditional generation shared the same information across modalities. Following [22–24,26,27], we considered the class label of the modalities as the shared information and used pretrained classifiers to extract the label information from the generated samples and compare it across modalities.

For **MNIST-SVHN**, **MHD**, and **POLYMNIST**, 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 a modality m with a subset of modalities A, the conditional generated sample $\hat{X}^m$ is fed to the modality-specific pretrained classifier $\mathbf{C}_m$. The predicted label class is compared to the ground truth label y_{X^A}, which is the label of the modalities in subset X^A. For N samples, the matching rate average establishes the coherence. For all the experiments, N was equal to the length of the test set.

$$Coherence(\hat{X}^m|X^A) = \frac{1}{N}\sum_{1}^{N} \mathbb{1}_{\{\mathbf{C}_m(\hat{X}^m)=y_{X^A}\}} \tag{A7}$$

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

The leave-one-out coherence is the conditional generation coherence using all possible subsets excluding the generated modality ($Coherence(\hat{X}^m|X^A)$ with $A = \{1, .., M\} \setminus m$). Due to the large number of modalities in **POLYMNIST**, similar to [23,24,27], we computed the

average **leave-one-out coherence** conditional coherence as a function of the subset size of the input modalities.

Due to the unavailability of labels in the **CUB** dataset, we used CLIP-S [56], a state-of-the art metric for image captioning evaluation.

Appendix C.2.2. Generation Quality

For each modality, we considered the following metrics:

- **RGB Images**: FID [54] is the state-of-the-art standard metric for evaluating the image generation quality of generative models.
- **Audio**: FAD [55] is a state-of-the-art standard metric for the evaluation of audio generation. FAD performs well in terms of robustness against noise, and is consistent with human judgments [65]. Similar to FID, the Fréchet distance is computed, except that VGGish (audio classifer model) embeddings are used instead.
- **Other modalities**: For other modality types, we derived the FMD (Fréchet Modality Distance), a similar metric to FID and FAD. We computed the **Fréchet distance** between the statistics retrieved from the activations of the modality-specific pretrained classifiers used for coherence evaluation. FMD was used to evaluate the generative quality of the MNIST modality on the **MNIST-SVHN** dataset and the image and trajectory modalities on the **MHD** dataset.

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

For **CUB**, we used 10,000 samples to evaluate the generation quality in terms of FID. In the remaining experiments, we used 5000 samples to evaluate the performance in terms of FID, FAD, or FMD.

Appendix D. Implementation Details

In this section, we report the implementation details for each benchmark. We used the same unified code base for all the baselines, relying on the *PyTorch* framework. The VAE implementation was adapted from the official code whenever available (MVAE, MMVAE and MOPOE, as in (https://github.com/thomassutter/MoPoE (accessed on 11 February 2024)), MVTCAE (https://github.com/gr8joo/MVTCAE (accessed on 11 February 2024)), and NEXUS (https://github.com/miguelsvasco/nexus_pytorch (accessed on 11 February 2024)). To ensure fairness, MLD and all VAE-based models used the same autoencoder architecture. We used the best hyperparameters suggested by the authors. Across all the datasets, we used the *Adam optimizer* [66] for training.

Appendix D.1. MLD

MLD used the same autoencoder architecture as for VAE-based models, except that the latter are deterministic autoencoders. The autoencoders were trained using the same reconstruction loss term as for the VAE-based models. Tables A5 and A6 summarize the hyperparameters used during the two phases of MLD training. Note that data augmentation was necessary for the image modality in the CUB dataset in order to overcome overfitting when training the deterministic autoencoder. For this, we used *TrivialAugmentWide* from the Torchvision library.

Table A5. MLD: hyperparameters used for the deterministic autoencoders.

Dataset	Modality	Latent Space	Batch Size	Lr	Epochs	Weight Decay
MNIST-SVHN	MNIST SVHN	16 64	128	1×10^{-3}	150	
MHD	Image Trajectory Sound	64 16 128	64	1×10^{-3}	500	
POLYMNIST	All modalities	160	128	1×10^{-3}	300	
CUB	Caption Image	32 64	128	1×10^{-3} 1×10^{-4}	500 300	1×10^{-6}
CelebAMask-HQ	Image Mask Attributes	256 128 32	64	1×10^{-3}	200	

Table A6. MLD: score network hyperparameters.

Dataset	d	Blocks	Width	Time Embed	Batch Size	Lr	Epochs
MNIST-SVHN	0.5	2	512	256	128		150
MHD	0.3	2	1024	512	128		3000
POLYMNIST	0.5	2	1536	512	256	1×10^{-4}	3000
CUB	0.7	2	1024	512	64		3000
CelebAMask-HQ	0.5	2	1536	512	64		3000

Appendix D.2. VAE-Based Models

For **MNIST-SVHN**, we followed [22,23] and used the same autoencoder architecture and pretrained classifier. The latent space size was set to 20, $\beta = 5.0$. For MVTCAE $\alpha = \frac{5}{6}$. For both modalities, the likelihood was estimated using the Laplace distribution. For NEXUS, we used the same modality latent space size as in MLD, the joint NEXUS latent space was set to 20, $\beta_i = 1.0$, and $\beta_c = 5.0$. We trained all the VAE-models for 150 epochs with a batch size of 256 and learning rate of 1×10^{-3}.

For **MHD**, we reused the autoencoder architecture and pretrained classifier from [26]. We adopted the hyperparameters from [26] to train the NEXUS model with the same settings while discarding the label modality. For the remaining VAE-based models, the latent space size was set to 128, $\beta = 1.0$, and $\alpha = \frac{5}{6}$ for MVTCAE. For all the modalities, Mean square error (MSE) was used to compute the reconstruction loss, similar to [26]. The models were trained for 600 epochs with a batch size of 128 and learning rate of 1×10^{-3}.

For **POLYMNIST**, we used the same autoencoder architecture and pretrained classifier used by [23,27]. We set the latent space size to 512, $\beta = 2.5$, and $\alpha = \frac{5}{6}$ for MVTCAE. For all the modalities, the likelihood was estimated using the Laplace distribution. For NEXUS, we used the same modality latent space size as in MLD, the joint NEXUS latent space was 64, $\beta_i = 1.0$, and $\beta_c = 2.5$. We trained all the models for 300 epochs with a batch size of 256 and learning rate of 1×10^{-3}.

For **CUB**, we used the same autoencoder architecture and implementation settings as in [24]. The Laplace and one-hot categorical distributions were used to estimate the likelihoods of the image and caption modalities, respectively. The latent space size was 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 used the same modality latent space sizes as in MLD, the joint NEXUS latent space was set to 64, $\beta_i = 1.0$, and $\beta_c = 1$. We trained all the models for 150 epochs with a batch size of 64. We used a learning rate of $5e-4$ for MVAE, MVTCAE, and MOPOE and 1×10^{-3} for the remaining models.

Finally, we note that in the official implementation of [23,27] on the **POLYMNIST** and **MNIST-SVHN** datasets, the classifiers were used for evaluation with dropout. In our implementation, we made sure to deactivate dropout during the evaluation step.

For **CelebAMask-HQ**, in our MLD experiments we used deterministic autoencoders instead of variational autoencoders [58].

Appendix D.3. MLD with Powerful Autoencoder

Here, we provide more detail about the CUB experiment using a more powerful autoencoder, denoted MLD* in Figure 5. We used an architecture similar to [10] adapted to (64×64) resolution images. We modified the autoencoder architecture to be deterministic and trained the model with a simple mean square error loss. We kept the same configuration as the CUB experiment described in the previous experiment on the same dataset, including the text autoencoder, score network, and hyperparameters. We performed further experiments with the same settings on 128×128 resolution images. We include the qualitative results in Figure A21.

Appendix D.4. Computation Resources

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

Appendix E. Additional Results

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

Appendix E.1. MNIST-SVHN

Appendix E.1.1. Self-Reconstruction

In Table A7, we report the results on *self-coherence*, which we use to support the arguments from Section 2. This metric is used to measure the loss of information due to latent collapse by showing the ability of all competing models to reconstruct an arbitrary modality given the same modality or a set thereof as an input. For our MLD model, self-reconstruction is done without using the diffusion model component; the modality is encoded using its deterministic encoder, and the decoder is fed the latent space to obtain the reconstruction.

We observe that the VAE-based models fail to reconstruct SVHN given SVHN. This is especially visible for the models based on the product-of-experts approach (MVAE and MVTCAE). In MLD, the deterministic autoencoders do not suffer from such weakness, and achieve the best overall performance.

Figure A9 shows the qualitative self-generation results. We remark that the digits in certain samples generated using VAE-based models differ from those in the input sample (for example, generation of the MNIST digit 3 in the case of MVAE and the SVHN digit 2 in the case of MVTCAE), indicating information loss due to latent collapse.

Table A7. Self-generation coherence and quality for **MNIST-SVHN** (M: MNIST, S: SVHN). The generation quality is measured in terms of FMD for MNIST and FID for SVHN. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Coherence (%↑)				Quality (↓)			
	M → M	**M,S → M**	**S → S**	**M,S → S**	**M → M**	**M,S → M**	**S → S**	**M,S → S**
MVAE	$86.92_{\pm 0.8}$	$88.03_{\pm 0.78}$	$40.62_{\pm 0.99}$	$68.01_{\pm 1.29}$	$10.75_{\pm 1.04}$	$10.79_{\pm 1.02}$	$60.22_{\pm 1.01}$	$59.0_{\pm 0.6}$
MMVAE	$87.22_{\pm 1.87}$	$77.35_{\pm 4.19}$	$67.31_{\pm 6.93}$	$39.44_{\pm 3.43}$	$12.15_{\pm 1.25}$	$20.24_{\pm 1.04}$	$58.1_{\pm 3.14}$	$171.42_{\pm 4.55}$
MOPOE	$89.95_{\pm 0.84}$	$91.71_{\pm 0.77}$	$67.26_{\pm 0.8}$	$\underline{83.58}_{\pm 0.44}$	$9.39_{\pm 0.76}$	$10.1_{\pm 0.73}$	$53.19_{\pm 1.06}$	$57.34_{\pm 1.35}$
NEXUS	$92.63_{\pm 0.45}$	$93.59_{\pm 0.4}$	$\underline{68.31}_{\pm 0.46}$	$83.13_{\pm 0.58}$	$4.92_{\pm 0.61}$	$5.16_{\pm 0.59}$	$85.67_{\pm 2.74}$	$97.86_{\pm 2.86}$
MVTCAE	$\underline{94.33}_{\pm 0.18}$	$\underline{95.18}_{\pm 0.19}$	$47.47_{\pm 0.76}$	$\mathbf{86.6}_{\pm 0.23}$	$\underline{4.67}_{\pm 0.35}$	$\underline{4.94}_{\pm 0.37}$	$\underline{52.29}_{\pm 1.17}$	$\underline{53.55}_{\pm 1.19}$
MLD	$\mathbf{96.73}_{\pm 0.0}$	$\mathbf{96.73}\pm 0.0$	$\mathbf{82.19}_{\pm 0.0}$	$82.19_{\pm 0.0}$	$\mathbf{2.25}_{\pm 0.03}$	$\mathbf{2.25}\pm 0.03$	$\mathbf{48.47}_{\pm 0.63}$	$\mathbf{48.47}_{\pm 0.63}$

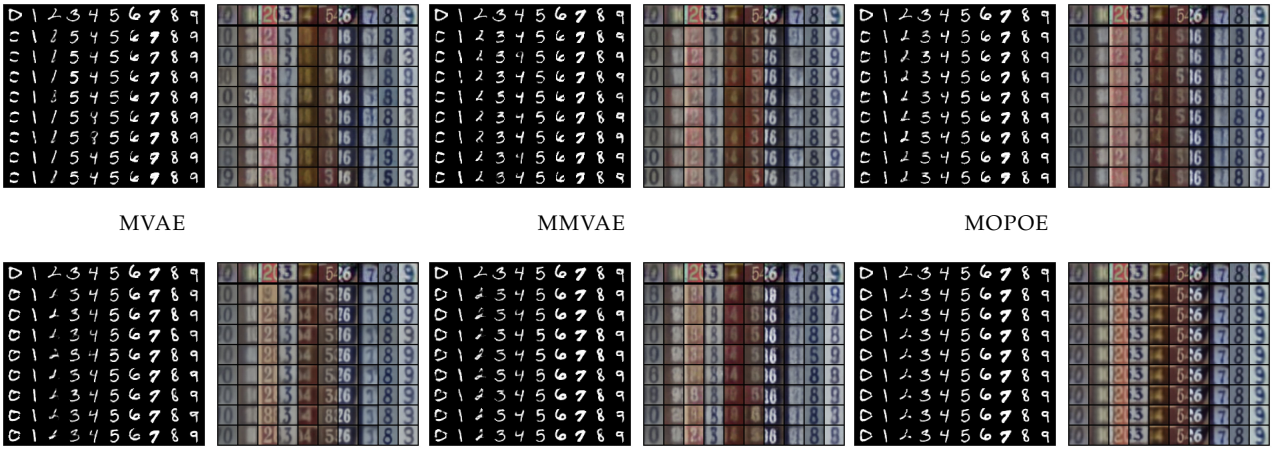


Figure A9. Self-generation qualitative results for **MNIST-SVHN**. For each model, we report MNIST-to-MNIST conditional generation on the left and SVHN-to-SVHN conditional generation on the right.

Appendix E.1.2. Detailed Results

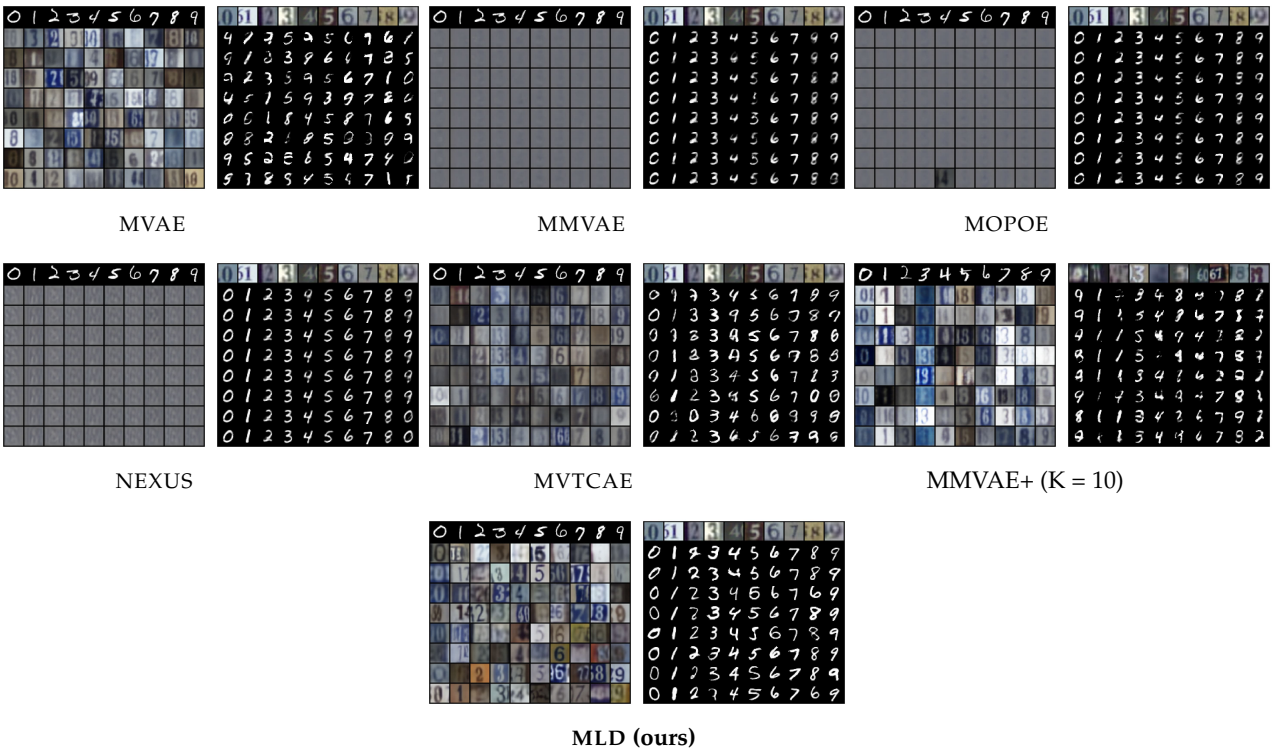


Figure A10. Additional qualitative results for **MNIST-SVHN**. For each model, we report MNIST-to-SVHN conditional generation on the left and SVHN-to-MNIST conditional generation on the right.

Table A8. Generative coherence for **MNIST-SVHN**. We report the detailed version of Table 1 with the standard deviation for five independent runs with different seeds. Bold and underlined numbers indicate the best and second best scores respectively.

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	$48.78_{\pm 1}$	$\underline{81.97}_{\pm 0.32}$	$49.78_{\pm 0.88}$	$12.98_{\pm 0.68}$	$\mathbf{52.92}_{\pm 1.39}$	$69.48_{\pm 1.64}$	$13.55_{\pm 0.8}$
MMVAE+	$17.64_{\pm 4.12}$	$13.23_{\pm 4.96}$	$29.69_{\pm 5.08}$	$26.60_{\pm 2.58}$	$121.77_{\pm 37.77}$	$240.90_{\pm 85.74}$	$35.11_{\pm 4.25}$
MMVAE+ (K = 10)	$41.59_{\pm 4.89}$	$55.3_{\pm 9.89}$	$56.41_{\pm 5.37}$	$19.05_{\pm 1.10}$	$67.13_{\pm 4.58}$	$75.9_{\pm 12.91}$	$18.16_{\pm 2.20}$
MLD	$\underline{85.22}_{\pm 0.5}$	$\mathbf{83.79}_{\pm 0.62}$	$\mathbf{79.13}_{\pm 0.38}$	$\underline{3.93}_{\pm 0.12}$	$\underline{56.36}_{\pm 1.63}$	$\mathbf{57.2}_{\pm 1.47}$	$\underline{3.67}_{\pm 0.14}$

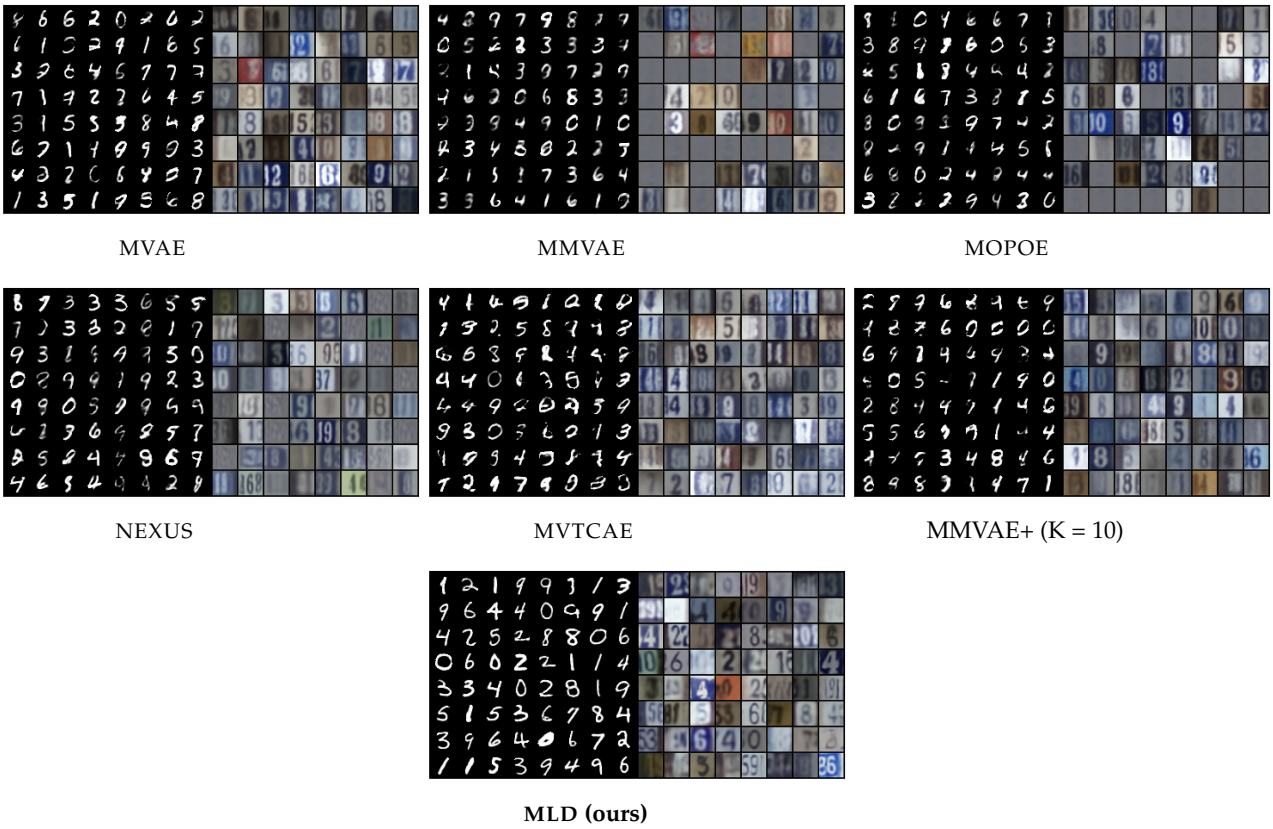


Figure A11. Qualitative results for **MNIST-SVHN** joint generation.

Appendix E.2. MHD

Table A9. Generative coherence for **MHD**. We report the detailed version of Table 2 with the standard deviation for five independent runs with different seeds. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Joint	I (Image)			T (Trajectory)			S (Sound)		
		T	S	T,S	I	S	I,S	I	T	I,T
MVAE	37.77$_{\pm 3.32}$	11.68$_{\pm 0.35}$	26.46$_{\pm 1.84}$	28.4$_{\pm 1.47}$	95.55$_{\pm 1.39}$	26.66$_{\pm 1.72}$	96.58$_{\pm 1.06}$	58.87$_{\pm 4.89}$	10.39$_{\pm 0.42}$	58.16$_{\pm 5.24}$
MMVAE	34.78$_{\pm 0.83}$	**99.7**$_{\pm 0.03}$	69.69$_{\pm 1.66}$	84.74$_{\pm 0.95}$	99.3$_{\pm 0.07}$	85.46$_{\pm 1.57}$	92.39$_{\pm 0.95}$	49.95$_{\pm 0.79}$	50.14$_{\pm 0.89}$	50.17$_{\pm 0.99}$
MOPOE	48.84$_{\pm 0.36}$	99.64$_{\pm 0.08}$	68.67$_{\pm 2.07}$	99.69$_{\pm 0.04}$	99.28$_{\pm 0.08}$	87.42$_{\pm 0.41}$	99.35$_{\pm 0.04}$	50.73$_{\pm 3.72}$	51.5$_{\pm 3.52}$	56.97$_{\pm 6.34}$
NEXUS	26.56$_{\pm 1.71}$	94.58$_{\pm 0.34}$	83.1$_{\pm 0.74}$	95.27$_{\pm 0.52}$	88.51$_{\pm 0.64}$	76.82$_{\pm 3.63}$	93.27$_{\pm 0.91}$	70.06$_{\pm 2.83}$	75.84$_{\pm 2.53}$	89.48$_{\pm 3.24}$
MVTCAE	42.28$_{\pm 1.12}$	99.54$_{\pm 0.07}$	72.05$_{\pm 0.95}$	99.63$_{\pm 0.05}$	99.22$_{\pm 0.08}$	72.03$_{\pm 0.48}$	99.39$_{\pm 0.02}$	92.58$_{\pm 0.47}$	93.07$_{\pm 0.36}$	94.78$_{\pm 0.25}$
MMVAE+	41.67$_{\pm 2.3}$	98.05$_{\pm 0.19}$	84.16$_{\pm 0.57}$	91.88$_{\pm}$	97.47$_{\pm 0.89}$	81.16$_{\pm 2.24}$	89.31$_{\pm 1.54}$	64.34$_{\pm 4.46}$	65.42$_{\pm 5.42}$	64.88$_{\pm 4.93}$
MMVAE+ (K = 10)	42.60$_{\pm 2.5}$	99.44$_{\pm 0.07}$	**89.75**$_{\pm 0.75}$	94.7$_{\pm 0.72}$	99.44$_{\pm 0.18}$	**89.58**$_{\pm 0.4}$	95.01$_{\pm 0.30}$	87.15$_{\pm 2.81}$	87.99$_{\pm 2.55}$	87.57$_{\pm 2.09}$
MLD	**98.34**$_{\pm 0.22}$	99.45$_{\pm 0.09}$	88.91$_{\pm 0.54}$	**99.88**$_{\pm 0.04}$	**99.58**$_{\pm 0.03}$	88.92$_{\pm 0.53}$	**99.91**$_{\pm 0.02}$	**97.63**$_{\pm 0.14}$	**97.7**$_{\pm 0.34}$	**98.01**$_{\pm 0.21}$

Table A10. Generative quality for **MHD**. We report the detailed version of Table 3 with the standard deviation for five independent runs with different seeds. Bold and underlined numbers indicate the best and second best scores respectively.

Models	I (Image)				T (Trajectory)				S (Sound)			
	Joint	T	S	T,S	Joint	I	S	I,S	Joint	I	T	I,T
MVAE	94.9$_{\pm 7.37}$	93.73$_{\pm 5.44}$	92.55$_{\pm 7.37}$	91.08$_{\pm 10.24}$	39.51$_{\pm 6.04}$	20.42$_{\pm 4.42}$	38.77$_{\pm 6.29}$	19.25$_{\pm 4.26}$	14.14$_{\pm 0.25}$	14.13$_{\pm 0.19}$	14.08$_{\pm 0.24}$	14.17$_{\pm 4.26}$
MMVAE	224.01$_{\pm 12.58}$	22.6$_{\pm 4.3}$	789.12$_{\pm 12.58}$	170.41$_{\pm 8.06}$	16.52$_{\pm 1.17}$	**0.5**$_{\pm 0.05}$	30.39$_{\pm 1.38}$	6.07$_{\pm 0.37}$	22.8$_{\pm 0.39}$	22.61$_{\pm 0.75}$	23.72$_{\pm 0.86}$	23.01$_{\pm 0.67}$
MOPOE	147.81$_{\pm 10.37}$	16.29$_{\pm 0.85}$	838.38$_{\pm 10.84}$	15.89$_{\pm 1.96}$	13.92$_{\pm 0.96}$	0.52$_{\pm 0.12}$	33.38$_{\pm 1.14}$	**0.53**$_{\pm 0.1}$	18.53$_{\pm 0.27}$	24.11$_{\pm 0.4}$	24.1$_{\pm 0.41}$	23.93$_{\pm 0.87}$
NEXUS	281.76$_{\pm 12.69}$	116.65$_{\pm 9.99}$	282.34$_{\pm 12.69}$	117.24$_{\pm 8.53}$	18.59$_{\pm 2.16}$	6.67$_{\pm 0.23}$	33.01$_{\pm 3.41}$	7.54$_{\pm 0.29}$	13.99$_{\pm 0.9}$	19.52$_{\pm 0.14}$	18.71$_{\pm 0.24}$	16.3$_{\pm 0.59}$
MVTCAE	121.85$_{\pm 3.44}$	5.34$_{\pm 0.33}$	54.57$_{\pm 7.79}$	3.16$_{\pm 0.26}$	19.49$_{\pm 0.67}$	0.62$_{\pm 0.1}$	13.65$_{\pm 1.24}$	0.75$_{\pm 0.13}$	15.88$_{\pm 0.19}$	14.22$_{\pm 0.27}$	14.02$_{\pm 0.14}$	13.96$_{\pm 0.28}$
MMVAE+	97.19$_{\pm 12.37}$	2.80$_{\pm 0.42}$	128.56$_{\pm 4.47}$	114.3$_{\pm 11.4}$	22.37$_{\pm 1.87}$	1.21$_{\pm 0.22}$	21.74$_{\pm 3.49}$	15.2$_{\pm 1.15}$	16.12$_{\pm 0.40}$	17.31$_{\pm 0.62}$	17.92$_{\pm 0.19}$	17.56$_{\pm 0.48}$
MMVAE+ (K = 10)	85.98$_{\pm 1.25}$	1.83$_{\pm 0.26}$	70.72$_{\pm 1.76}$	62.43$_{\pm 3.4}$	21.10$_{\pm 1.25}$	1.38$_{\pm 0.34}$	8.52$_{\pm 0.79}$	7.22$_{\pm 1.6}$	14.58$_{\pm 0.47}$	14.33$_{\pm 0.51}$	14.34$_{\pm 0.42}$	14.32$_{\pm 0.6}$
MLD (ours)	**7.98**$_{\pm 1.41}$	**1.7**$_{\pm 0.14}$	**4.54**$_{\pm 0.45}$	**1.84**$_{\pm 0.27}$	**3.18**$_{\pm 0.18}$	0.83$_{\pm 0.03}$	**2.07**$_{\pm 0.26}$	0.6$_{\pm 0.05}$	**2.39**$_{\pm 0.1}$	**2.31**$_{\pm 0.07}$	**2.33**$_{\pm 0.11}$	**2.29**$_{\pm 0.06}$

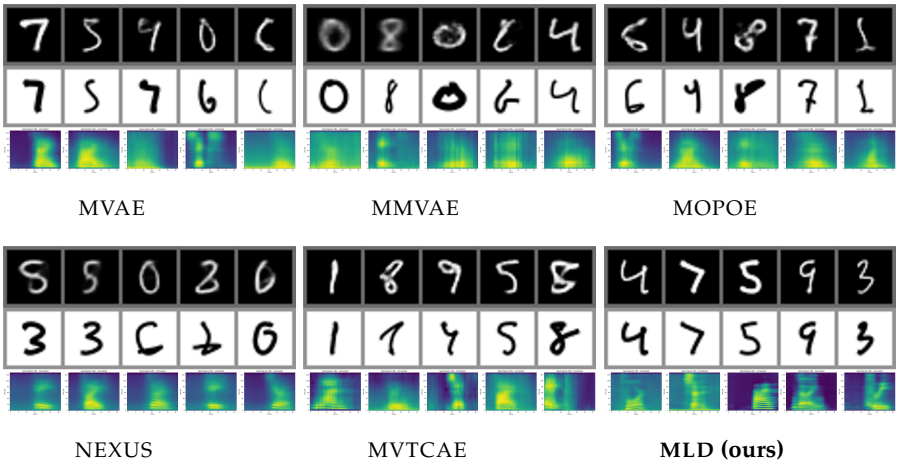


Figure A12. Joint generation qualitative results for **MHD**. The three modalities were randomly generated simultaneously. **Top row**: image; **Middle row**: trajectory vector converted into image; **Bottom row**: sound mel-spectrogram).

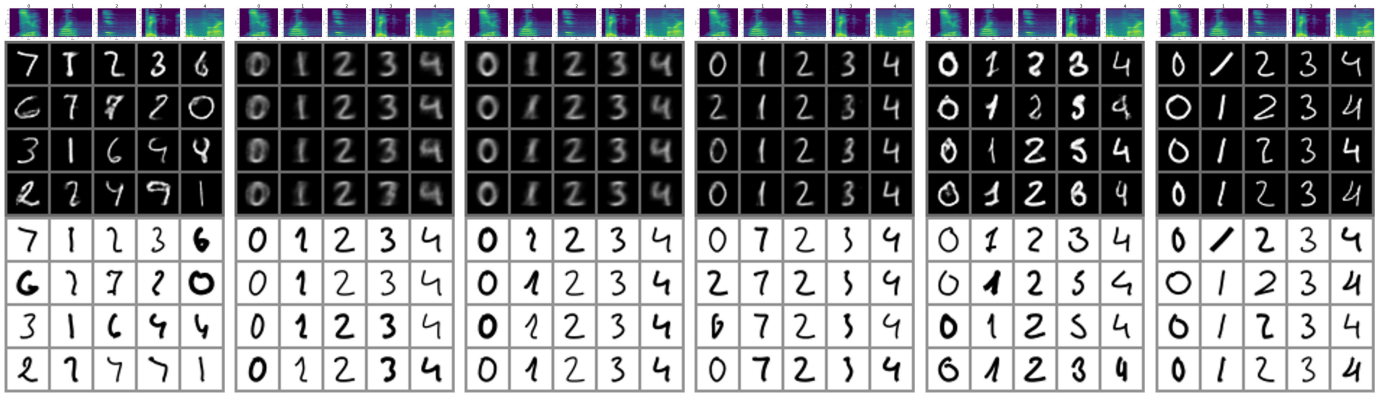


Figure A13. Sound-to-image and trajectory conditional generation qualitative results for **MHD**. For each model, the **Top row** reports the sound mel-spectrograms of the digits {0,1,2,3,4} from left to right and the **Lower rows** report the generated image and trajectory samples.

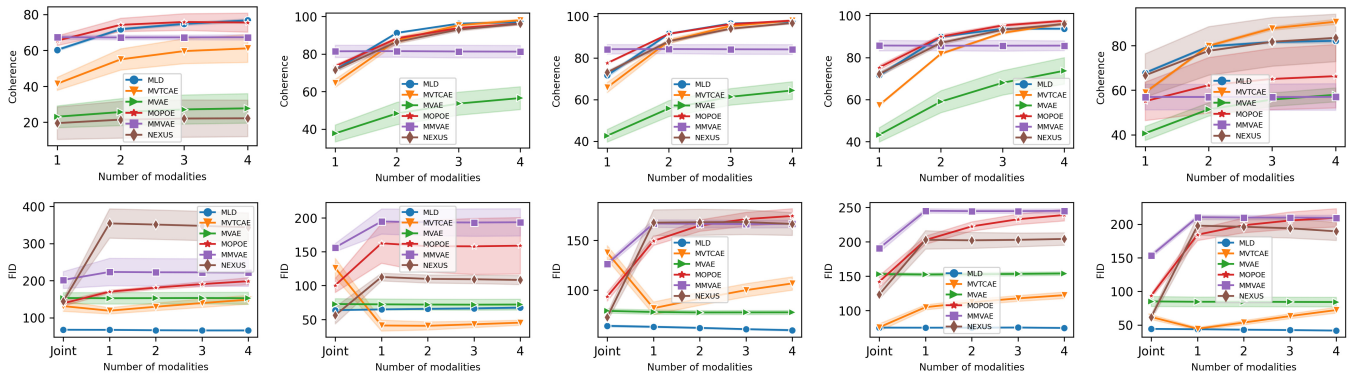
Appendix E.3. POLYMNIST

Table A11. Generation coherence (%) for **POLYMNIST** (higher is better) used for the plots in Figure 4 and Figure A5. We report the average *leave-one-out coherence* as a function of the number of observed modalities. *Joint* refers to random generation of the five modalities simultaneously. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Coherence (%↑)				
	Joint	1	2	3	4
MVAE	$4.0_{\pm 1.49}$	$37.51_{\pm 3.16}$	$48.06_{\pm 3.55}$	$53.19_{\pm 3.37}$	$56.09_{\pm 3.31}$
MMVAE	$25.8_{\pm 1.43}$	$\mathbf{75.15}_{\pm 2.54}$	$75.14_{\pm 2.47}$	$75.09_{\pm 2.6}$	$75.09_{\pm 2.58}$
MOPOE	$17.32_{\pm 2.47}$	$\underline{69.37}_{\pm 1.85}$	$\underline{81.29}_{\pm 2.34}$	$85.26_{\pm 2.36}$	$86.7_{\pm 2.39}$
NEXUS	$18.24_{\pm 0.89}$	$60.61_{\pm 2.51}$	$72.14_{\pm 2.79}$	$76.81_{\pm 2.75}$	$78.92_{\pm 2.64}$
MVTCAE	$0.21_{\pm 0.05}$	$57.66_{\pm 1.06}$	$78.44_{\pm 1.31}$	$\underline{85.97}_{\pm 1.43}$	$\underline{88.81}_{\pm 1.49}$
MMVAE+	$26.28_{\pm 2.19}$	$54.74_{\pm 0.5}$	$54.06_{\pm 0.33}$	$55.2_{\pm 1.32}$	$53.17_{\pm 0.75}$
MMVAE+ (K = 10)	$14.53_{\pm 4.94}$	$58.93_{\pm 6.3}$	$59.42_{\pm 8.8}$	$60.77_{\pm 8.03}$	$58.24_{\pm 7.42}$
MLD IN-PAINT	$\underline{51.65}_{\pm 1.16}$	$52.85_{\pm 0.23}$	$77.65_{\pm 0.24}$	$85.66_{\pm 0.43}$	$87.29_{\pm 0.29}$
MLD UNI	$48.79_{\pm 0.43}$	$65.12_{\pm 0.7}$	$79.52_{\pm 0.8}$	$82.03_{\pm 1.19}$	$81.86_{\pm 2.09}$
MLD	$\mathbf{56.23}_{\pm 0.52}$	$68.58_{\pm 0.72}$	$\mathbf{84.87}_{\pm 0.19}$	$\mathbf{88.56}_{\pm 0.12}$	$\mathbf{89.43}_{\pm 0.27}$

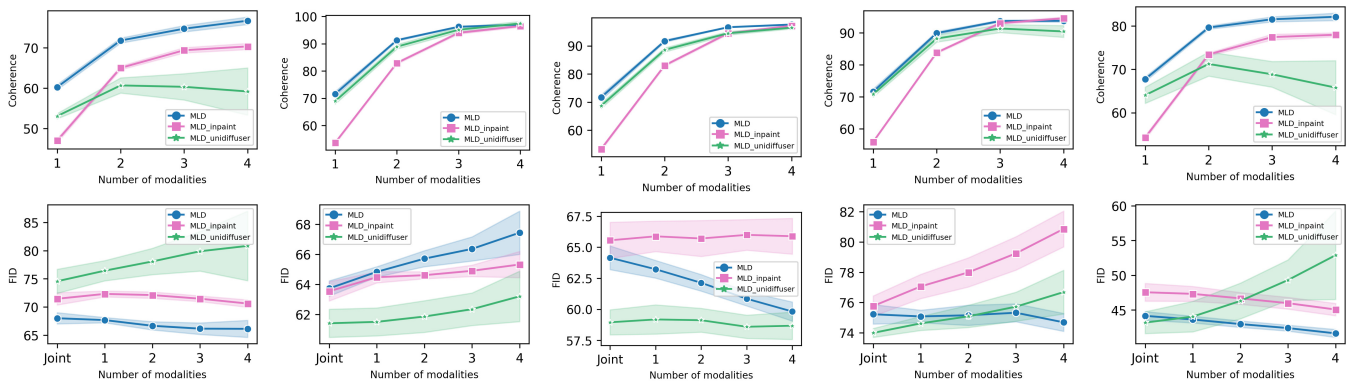
Table A12. Generation quality (FID ↓) for **POLYMNIST** (lower is better) used for the plots in Figures 4 and A5. Similar to Table A11, we report the average *leave-one-out* FID as a function of the number of observed modalities. *Joint* refers to random generation of the five modalities simultaneously. Bold and underlined numbers indicate the best and second best scores respectively.

Models	Quality (↓)				
	Joint	1	2	3	4
MVAE	$108.74_{\pm 2.73}$	$108.06_{\pm 2.79}$	$108.05_{\pm 2.73}$	$108.14_{\pm 2.71}$	$108.18_{\pm 2.85}$
MMVAE	$165.74_{\pm 5.4}$	$208.16_{\pm 10.41}$	$207.5_{\pm 10.57}$	$207.35_{\pm 10.59}$	$207.38_{\pm 10.58}$
MOPOE	$113.77_{\pm 1.62}$	$173.87_{\pm 7.34}$	$185.06_{\pm 10.21}$	$191.72_{\pm 11.26}$	$196.17_{\pm 11.66}$
NEXUS	$91.66_{\pm 2.93}$	$207.14_{\pm 7.71}$	$205.54_{\pm 8.6}$	$204.46_{\pm 9.08}$	$202.43_{\pm 9.49}$
MVTCAE	$106.55_{\pm 3.83}$	$78.3_{\pm 2.35}$	$85.55_{\pm 2.51}$	$92.73_{\pm 2.65}$	$99.13_{\pm 2.72}$
MMVAE+	$168.88_{\pm 0.12}$	$165.67_{\pm 0.14}$	$166.5_{\pm 0.18}$	$165.53_{\pm 0.55}$	$165.3_{\pm 0.33}$
MMVAE+ (K = 10)	$156.55_{\pm 3.58}$	$154.42_{\pm 2.73}$	$153.1_{\pm 3.01}$	$153.06_{\pm 2.88}$	$154.9_{\pm 2.9}$
MLD IN-PAINT	$64.78_{\pm 0.33}$	$65.41_{\pm 0.43}$	$65.42_{\pm 0.41}$	$65.52_{\pm 0.46}$	$\underline{65.55}_{\pm 0.46}$
MLD UNI	$\mathbf{62.42}_{\pm 0.62}$	$\underline{63.16}_{\pm 0.81}$	$\underline{64.09}_{\pm 1.15}$	$\underline{65.17}_{\pm 1.46}$	$66.46_{\pm 2.18}$
MLD	$\underline{63.05}_{\pm 0.26}$	$\mathbf{62.89}_{\pm 0.2}$	$\mathbf{62.53}_{\pm 0.21}$	$\mathbf{62.22}_{\pm 0.39}$	$\mathbf{61.94}_{\pm 0.65}$



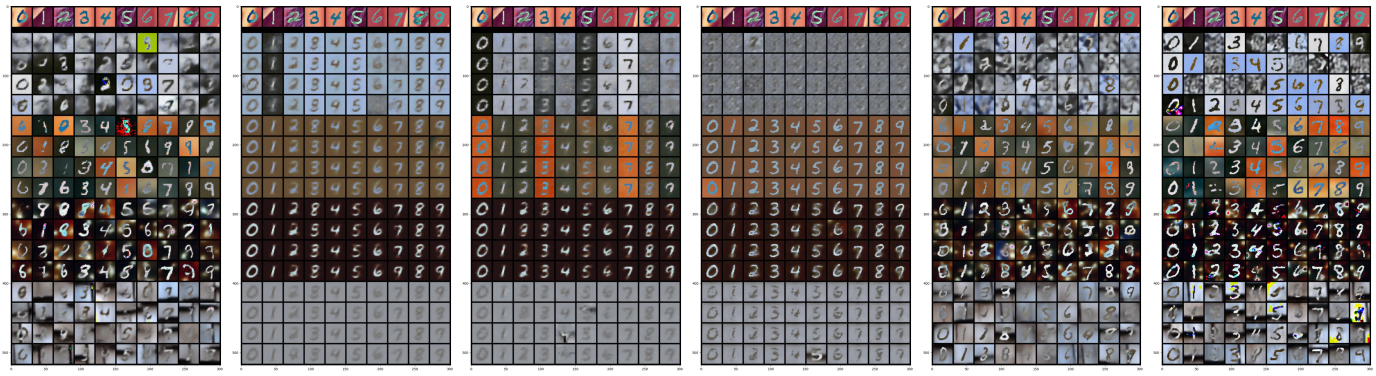
(**a**) X^0 (**b**) X^1 (**c**) X^2 (**d**) X^3 (**e**) X^4

Figure A14. Top: Generation coherence (%) for **POLYMNIST** (higher is better). **Bottom:** Generation quality (FID) (lower is better). We report the average *leave-one-out* performance as a function of the number of observed modalities for each modality X^i. *Joint* refers to random generation of the five modalities simultaneously.



(**a**) X^0 (**b**) X^1 (**c**) X^2 (**d**) X^3 (**e**) X^4

Figure A15. Top: Generation coherence (%) for **POLYMNIST** (higher is better).**Bottom:** Generation quality (FID) (lower is better). We report the average *leave-one-out* performance as a function of the number of observed modalities for each modality X^i. *Joint* refers to random generation of the five modalities simultaneously.



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

Figure A16. Conditional generation qualitative results for **POLYMNIST**. Modality X^2 (first row) is used as the condition to generate the four remaining modalities (the rows below).

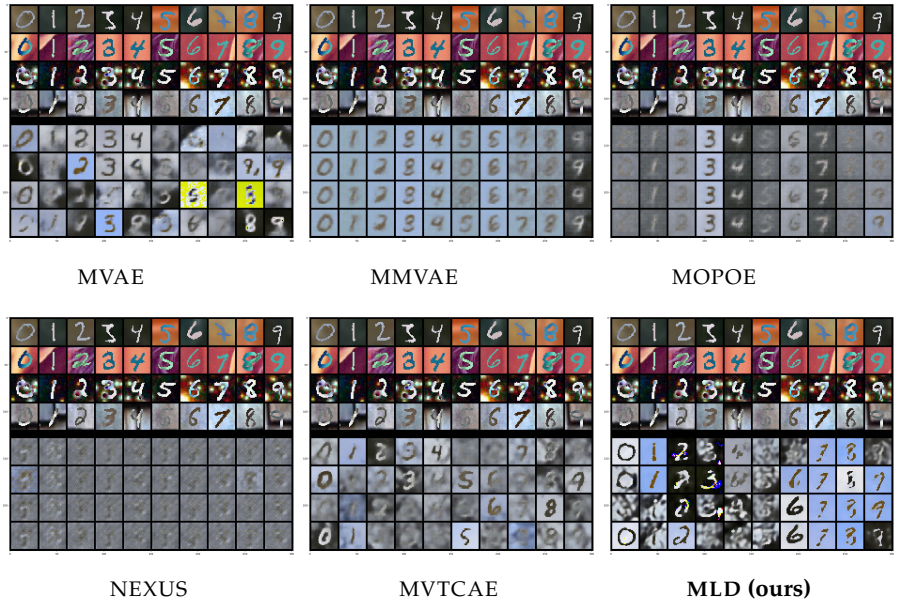


Figure A17. Conditional generation qualitative results for **POLYMNIST**. The subset of modalities X^1, X^2, X^3, X^4 (first four rows) are used as the condition to generate modality X^0 (the rows below).

Additional Experiments with the Architecture from [29]

In our experiments on POLYMNIST, we used the same architecture as in [23,27] in order to ensure a fair settings for all the baselines. In [29], the experiments on POLYMNIST were conducted using a different autoencoder architecture based on Resnet instead of a sequence of autoencoder-based convolutional layers. In this section, we investigate the performance of MMVAE+ and our MLD using this architecture. For MMVAE+, we kept the same settings as in [29], including the autoencoder architecture, latent size, and importance sampling K = 10 with doubly reparameterized gradient estimator (DReG). For MLD, we used the same autoencoder architecture with a latent size equal to 160. In Figure A18, can be observed that while the new autoencoder architecture enhances the performance of MMVAE+, the performance our MLD is improved as well. Similar to the previous results, MLD simultaneously achieves the best generative coherence and the best quality.

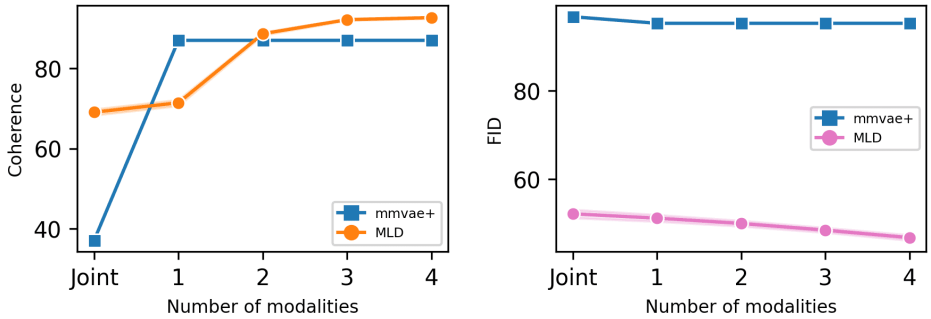


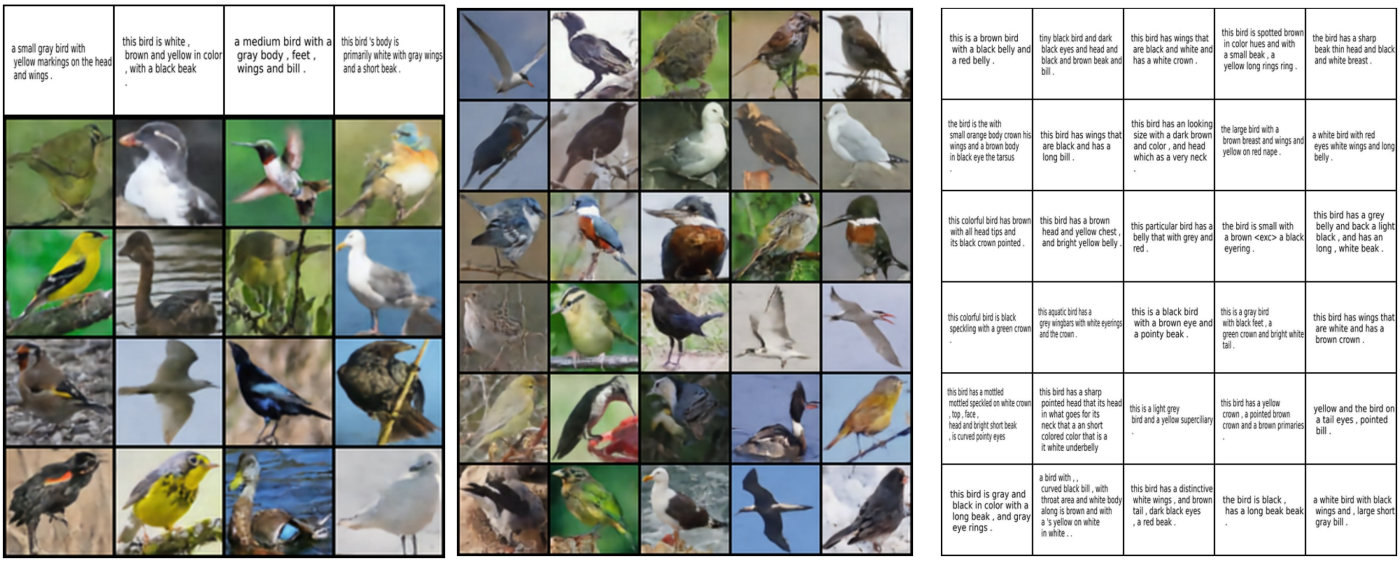
Figure A18. Results for the POLYMNIST dataset. Left: Comparison of the generative coherence (↑) and quality in terms of FID (↓) as a function of the number of inputs.

Appendix E.4. CUB

Table A13. Generation coherence (CLIP-S: higher is better) and quality (FID: ↓ lower is better) for the CUB dataset. **MLD*** denotes the version of our method using a more powerful image autoencoder. Bold numbers indicate the best scores.

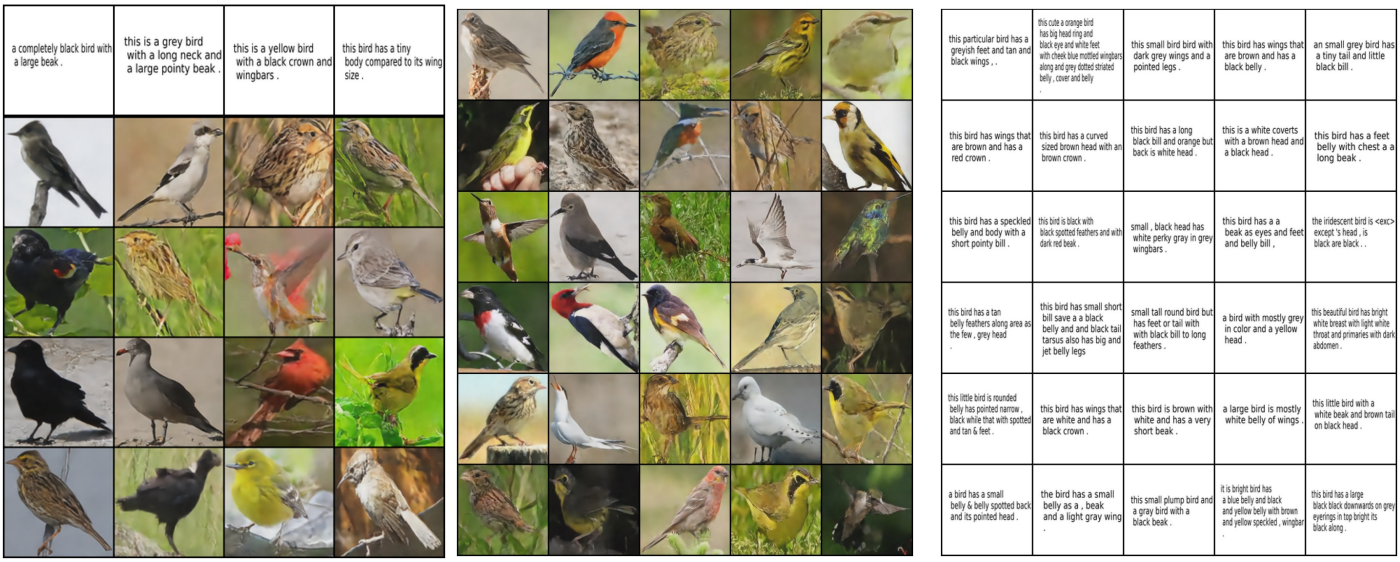
Models	Coherence (↑)			Quality (↓)	
	Joint	Image → Caption	Caption → Image	Joint → Image	Caption → Image
MVAE	0.66	**0.70**	0.64	158.91	158.88
MMVAE	0.66	0.69	0.62	277.8	212.57
MOPOE	0.64	0.68	0.55	279.78	179.04
NEXUS	0.65	0.69	0.59	147.96	262.9
MVTCAE	0.65	**0.70**	0.65	155.75	168.17
MMVAE+	0.61	0.68	0.65	188.63	247.44
MMVAE+ (K=10)	0.63	0.68	0.62	172.21	178.88
MLD IN-PAINT	**0.69**	0.69	0.68	69.16	68.33
MLD UNI	**0.69**	0.69	**0.69**	64.09	**61.92**
MLD	**0.69**	0.69	**0.69**	**63.47**	62.62
MLD*	**0.70**	0.69	**0.69**	**22.19**	**22.50**

Figure A19. Qualitative results for joint generation on the **CUB** dataset (Better viewed zoomed).



(**a**) Conditional generation. (**b**) Joint generation.

Figure A20. Qualitative results of **MLD*** on the **CUB** dataset with powerful image autoencoder (Better viewed zoomed).



(**a**) Conditional Generation (**b**) Joint generation

Figure A21. Qualitative results of **MLD*** on the **CUB** dataset with 128 $\times$ 128 resolution images and powerful image autoencoder (Better viewed zoomed).

Appendix E.5. CelebAMask-HQ

In this section, we present additional experiments on the CelebAMask-HQ dataset [58].

Figure A22. (Mask → Image) Conditional generation of **MLD** on the CelebAMask-HQ dataset.

References

1. He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; QI, X. Is Synthetic Data from Generative Models Ready for Image Recognition? In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
2. Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; Fleet, D.J. Synthetic Data from Diffusion Models Improves ImageNet Classification. *arXiv* **2023**, arXiv:2304.08466.
3. Sariyildiz, M.B.; Alahari, K.; Larlus, D.; Kalantidis, Y. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. *arXiv* **2023**, arXiv:2212.08420.
4. Antelmi, L.; Ayache, N.; Robert, P.; Lorenzi, M. Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 302–311.
5. Da Silva–Filarder, M.; Ancora, A.; Filippone, M.; Michiardi, P. Multimodal Variational Autoencoders for Sensor Fusion and Cross Generation. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Virtual, 13–16 December 2021; pp. 1069–1076. [CrossRef]
6. Zhang, Y.; Peng, C.; Wang, Q.; Song, D.; Li, K.; Zhou, S.K. Unified Multi-Modal Image Synthesis for Missing Modality Imputation. *arXiv* **2023**, arXiv:2304.05340.
7. Tran, L.; Liu, X.; Zhou, J.; Jin, R. Missing Modalities Imputation via Cascaded Residual Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
8. Huang, X.; Mallya, A.; Wang, T.C.; Liu, M.Y. Multimodal Conditional Image Synthesis With Product-of-Experts GANs. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Part XVI, Tel Aviv, Israel, 23–27 October 2022; pp. 91–109. [CrossRef]
9. Lee, S.; Ha, J.; Kim, G. Harmonizing Maximum Likelihood with GANs for Multimodal Conditional Generation. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
10. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
11. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Gontijo-Lopes, R.; Ayan, B.K.; Salimans, T.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Proceedings of the NeurIPS 2022, New Orleans, LA, USA, 28 November–9 December 2022.
12. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125.
13. Tao, M.; Tang, H.; Wu, F.; Jing, X.Y.; Bao, B.K.; Xu, C. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. *arXiv* **2022**, arXiv:2008.05865.

14. Wu, F.; Liu, L.; Hao, F.; He, F.; Cheng, J. 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), New Orleans, LA, USA, 18–24 June 2022; pp. 18113–18122.
15. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv* **2022**, arXiv:2112.10741.
16. Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.H.; Murphy, K.; Freeman, W.T.; Rubinstein, M.; et al. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv* **2023**, arXiv:2301.00704.
17. Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S.W.; Fidler, S.; Kreis, K. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. *arXiv* **2023**, arXiv:2304.08818.
18. Hong, W.; Ding, M.; Zheng, W.; Liu, X.; Tang, J. CogVideo: Large-Scale Pretraining for Text-to-Video Generation via Transformers. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
19. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv* **2022**, arXiv:2209.14792.
20. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
21. Wu, M.; Goodman, N. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In Proceedings of the NeurIPS 2018, Montreal, QC, Canada, 2–8 December 2018.
22. Shi, Y.; N, S.; Paige, B.; Torr, P. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In Proceedings of the NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019.
23. Sutter, T.M.; Daunhawer, I.; Vogt, J.E. Generalized Multimodal ELBO. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
24. Daunhawer, I.; Sutter, T.M.; Chin-Cheong, K.; Palumbo, E.; Vogt, J.E. On the Limitations of Multimodal VAEs. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
25. Shi, Y.; Paige, B.; Torr, P.; N, S. Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
26. Vasco, M.; Yin, H.; Melo, F.S.; Paiva, A. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Netw.* **2022**, *146*, 238–255. [CrossRef]
27. Hwang, H.; Kim, G.H.; Hong, S.; Kim, K.E. Multi-View Representation Learning via Total Correlation Objective. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12194–12207.
28. Sutter, T.M.; Daunhawer, I.; Vogt, J.E. Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6100–6110.
29. Palumbo, E.; Daunhawer, I.; Vogt, J.E. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
30. Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R.A.; Murphy, K. Fixing a broken ELBO. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 159–168.
31. Dieng, A.B.; Kim, Y.; Rush, A.M.; Blei, D.M. Avoiding latent variable collapse with generative skip models. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Naha, Okinawa, Japan, 16–18 April 2019.
32. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
33. Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N.J.; Jin, Q.; Guo, B. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. *arXiv* **2023**, arXiv:2212.09478.
34. Hu, M.; Zheng, C.; Yang, Z.; Cham, T.J.; Zheng, H.; Wang, C.; Tao, D.; Suganthan, P.N. Unified Discrete Diffusion for Simultaneous Vision-Language Generation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
35. Bao, F.; Nie, S.; Xue, K.; Li, C.; Pu, S.; Wang, Y.; Yue, G.; Cao, Y.; Su, H.; Zhu, J. One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. *arXiv* **2023**, arXiv:2303.06555.
36. Wesego, D.; Rooshenas, A. Score-Based Multimodal Autoencoders. *arXiv* **2023**, arXiv:2303.06555.
37. Asperti, A.; Trentin, M. Balancing Reconstruction Error and Kullback-Leibler Divergence in Variational Autoencoders. *IEEE Access* **2020**, *8*, 199440–199448. [CrossRef]
38. Javaloy, A.; Meghdadi, M.; Valera, I. Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022.
39. Loaiza-Ganem, G.; Ross, B.L.; Cresswell, J.C.; Caterini, A.L. Diagnosing and Fixing Manifold Overfitting in Deep Generative Models. *arXiv* **2022**, arXiv:2204.07172.
40. Tran, B.H.; Rossi, S.; Milios, D.; Michiardi, P.; Bonilla, E.V.; Filippone, M. Model selection for bayesian autoencoders. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 19730–19742.
41. Dai, B.; Wipf, D. Diagnosing and enhancing VAE models. *arXiv* **2019**, arXiv:1903.05789.
42. Vahdat, A.; Kreis, K.; Kautz, J. Score-based Generative Modeling in Latent Space. In Proceedings of the NeurIPS 2021, Virtual, 6–14 December 2021.

43. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015.
44. Oksendal, B. *Stochastic Differential Equations: An Introduction with Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
45. Song, Y.; Durkan, C.; Murray, I.; Ermon, S. Maximum likelihood training of score-based diffusion models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1415–1428.
46. Anderson, B.D. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **1982**, *12*, 313–326. [CrossRef]
47. Franzese, G.; Rossi, S.; Yang, L.; Finamore, A.; Rossi, D.; Filippone, M.; Michiardi, P. How Much Is Enough? A Study on Diffusion Times in Score-Based Generative Models. *Entropy* **2023**, *25*, 633. [CrossRef] [PubMed]
48. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598.
49. Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; Bansal, M. Any-to-Any Generation via Composable Diffusion. *arXiv* **2023**, arXiv:2305.11846.
50. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.S. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv* **2023**, arXiv:2309.05519.
51. Xie, S.M.; Raghunathan, A.; Liang, P.; Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
52. Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *arXiv* **2022**, arXiv:2202.12837.
53. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11461–11471.
54. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
55. Kilgour, K.; Zuluaga, M.; Roblek, D.; Sharifi, M. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2350–2354. [CrossRef]
56. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv* **2021**, arXiv:2104.08718.
57. Kim, J.H.; Kim, Y.; Lee, J.; Yoo, K.M.; Lee, S.W. Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. *arXiv* **2022**, arXiv:2205.13445.
58. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5548–5557.
59. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
60. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
61. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
62. Dupont, E.; Kim, H.; Eslami, S.M.A.; Rezende, D.J.; Rosenbaum, D. From data to functa: Your data point is a function and you can treat it like one. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022.
63. Song, Y.; Ermon, S. Improved Techniques for Training Score-Based Generative Models. In Proceedings of the NeurIPS 2020, Virtual, 6–12 December 2020.
64. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
65. Vinay, A.; Lerch, A. Evaluating generative audio systems and their metrics. *arXiv* **2022**, arXiv:2209.00130.
66. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.