Multi-Component VAE with Gaussian Markov Random Field

Fouad Oubari[1,2] **Mohamed El Baha**[3] **Raphaël Meunier**[2]
Rodrigue Décatoire[2] **Mathilde Mougeot**[1,3]
[1]Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli [2]Michelin [3]ENSIIE
oubarifouad@gmail.com elbahamohamed@gmail.com raphael.meunier@michelin.com
rodrigue.decatoire@michelin.com mathilde.mougeot@ens-paris-saclay.fr

Abstract

Multi-component datasets with intricate dependencies challenge current generative modeling techniques. Existing Multi-component Variational AutoEncoders rely on simplified aggregation strategies that compromise structural coherence across generated components. We introduce the Gaussian Markov Random Field Multi-Component Variational AutoEncoder, embedding Gaussian Markov Random Fields into both prior and posterior distributions to explicitly model cross-component relationships, enabling richer representation and faithful reproduction of complex interactions. Empirically, our model achieves state-of-the-art performance on a synthetic Copula dataset designed for intricate component relationships, competitive results on PolyMNIST, and significantly enhanced structural coherence on the real-world BIKED dataset.[1]

1 INTRODUCTION

Multi-component datasets, ranging from medical imaging (CT-MRI pairs) (Puhr-Westerheide et al., 2019; Rahimi et al., 2022) to finance (multiple correlated markets) (Xie et al., 2024; Lee and Yoo, 2020) and industrial design (structured assemblies) (Cobb et al., 2023; Oubari et al., 2024) are prevalent in real-world applications. Generating such data requires models that capture both component-wise intricacies and cross-component interactions.

Although various generative approaches exist for multimodal data (Wu and Goodman, 2018; Shi et al., 2019), many rely on simplified aggregation schemes that overlook nuanced inter-component factors. Hence, we investigate Markov Random Fields (MRFs) as a structured way to encode the relationships more explicitly. By decomposing the latent distribution into terms that reflect how components relate to each other (Koller and Friedman, 2009), MRFs allow more nuanced interactions to be embedded, potentially leading to generations that better preserve global consistency. This is particularly relevant in conditional multi-component generation, where observing one component should constrain the plausible configurations of the others, and where the coherence of the assembled system is a central requirement.

1.1 Contributions

We propose the Gaussian Markov Random Field Multi-Component VAE (GMRF MCVAE) with the following contributions:

(i) **Gaussian MRF Multi-Component VAE:** Novel architecture embedding GMRFs in both prior and posterior for richer inter-component correlations. (ii) **Practical Covariance Construction:** Blockwise assembly guaranteeing symmetric positive definiteness with closed-form conditional generation. (iii) **Robust Empirical Validation:** State-of-the-art on synthetic Copula data, competitive results on PolyMNIST, and improved structural coherence on real-world BIKED, with competitive FID.

2 RELATED WORK

2.1 Multi-Component VAEs

The field of multi-component generative models has seen substantial growth recently. Within this domain,

[1]Code available at `https://github.com/foubari/gmrf_mcvae`.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

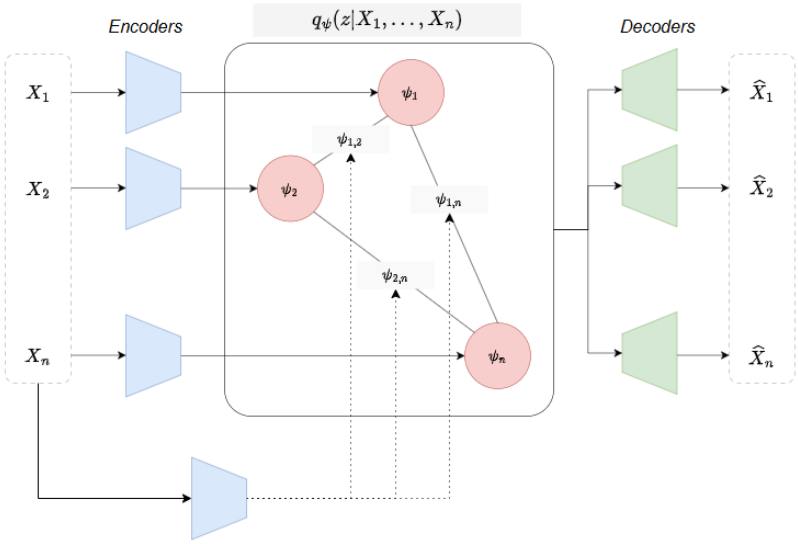

Figure 1: A general MRF-based Multi-Component VAE: each component is assigned its own encoder-decoder pair, where the encoder learns unary potentials ψ_i. A global encoder models pairwise potentials $\psi_{i,j}$ among components. Sampling $\mathbf{z}$ from this MRF-based latent space captures cross-component relationships. In practice, we adopt a Gaussian assumption for computational simplicity.

VAE-based models have distinguished themselves due to their rapid and tractable sampling capabilities (Vahdat and Kautz, 2020), as well as their robust generalization performance (Mbacke et al., 2024). The essence of multi-component generation lies in its ability to learn a joint latent representation from multiple data components, encapsulating a unified distribution. Traditional MCVAE frameworks typically adopt a structure with separate encoder/decoder pairs for each component, coupled with an aggregation mechanism to encode a cohesive joint representation across all components. A variety of methodologies have been introduced to synthesize these distributions within the latent space.

A seminal approach by Wu and Goodman (2018) suggests that the joint latent posterior can be effectively approximated through using the Product of Experts (PoE) assumption. This strategy facilitates the generation of cross components at inference time without requiring an additional inference network or a multi-stage training process, marking a significant advancement over previous methodologies (Suzuki et al., 2016; Vedantam et al., 2017). However, this approach implicitly relies on the assumption that the posterior distribution can be approximated by factorizing distributions. This assumption presupposes independence among components, which may not be true. This assumption overlooks the complex inter-component relationships intrinsic to the data, potentially limiting the model's ability to fully capture the richness of multi-component interactions.

An alternative framework proposed by Shi et al. (2019) employs a Mixture of Experts (MoE) strategy for aggregating marginal posteriors. This method stands in contrast to the approach used in the MVAE (Wu and Goodman, 2018), which, according to the authors, is susceptible to a 'veto phenomenon' a scenario where an exceedingly low marginal posterior density significantly diminishes the joint posterior density. In contrast, the MoE paradigm mitigates the risk associated with overly confident experts by adopting a voting mechanism among experts, thereby distributing its density among all contributing experts. However, a critique by Palumbo et al. (2023) highlights a fundamental limitation of the MMVAE approach: it tends to average the contribution of each component. Given that the model employs each component-specific encoder to reconstruct all other components, the resultant encoding is biased towards information that is common across all components. This bias towards commonality potentially undermines the model's ability to capture and represent the diversity inherent in multi-component datasets.

The mix of experts' products (MoPoE) framework (Sutter et al., 2021) refines and generalizes the aggregation strategies of PoE and MoE, combining the precise joint posterior approximation of PoE with the improved learning of component-specific posteriors by

MoE. The MoPoE model is designed to enhance multi-component learning by integrating these traits. Despite its conceptual advancements, the MoPoE model introduces a computational challenge due to its training strategy. It necessitates the evaluation of all conceivable component subsets, which equates to $2^M - 1$ training configurations for M components. This comprehensive strategy, while beneficial for robust learning across varied component combinations, leads to an exponential increase in computational requirements relative to the number of components. This aspect marks a significant limitation, especially for applications involving a large number of components.

To mitigate the averaging problem observed in mixture-based models, several studies (Sutter et al., 2020; Palumbo et al., 2023) have adopted component-specific latent spaces. Specifically, Palumbo et al. (2023) identifies a 'shortcut' phenomenon, characterized by information predominantly circulating within component-specific subspaces. To address this, an enhancement of Shi et al. (2019)'s model incorporates component-specific latent spaces designed exclusively for self-reconstruction. This strategy prevents the 'shortcut' by using a shared latent space to aggregate and a component-specific space to reconstruct unobserved components, ensuring that only joint information is retained in the shared space. Despite this advancement over prior approaches by resolving the shortcut dilemma, the outlined method introduces a training procedure that encompasses both reconstruction and cross-reconstruction tasks for each component pairing, leading to a computational requirement of M^2 forward passes for M components.

2.2 Markov Random Fields

Undirected Graphical Models, also called Markov Random Fields (Wainwright et al., 2008; Koller and Friedman, 2009; Murphy, 2012), were introduced to probability theory as a way to extend Markov processes from a temporal framework to a spatial one; they represent a stochastic process that has its origins in statistical physics (Kindermann and Snell, 1980). Graphical models, including MRF, are notoriously hard to train due to the intractability of the partition function. This has led to numerous studies (Carreira-Perpinan and Hinton, 2005; Vuffray et al., 2020; Bach and Jordan, 2002; Tan et al., 2014; Welling and Sutton, 2005) aimed at developing more efficient methods for learning graphical models, including MRF.

2.3 MRF in Machine Learning

Markov Random Fields have predominantly been used in image processing tasks such as image deblurring (Perez et al., 1998), completion, texture synthesis, and image inpainting (Komodakis and Tziritas, 2007), as well as segmentation (Krähenbühl and Koltun, 2011; Bello, 1994). However, recent advancements in more efficient methodologies have led to a decline in the use of MRF, due to the relative complexity involved in their learning processes.

To the best of our knowledge, there are limited instances where Markov Random Fields have been integrated within generative neural networks. Among these, Johnson et al. (2016) introduced the Structured Variational AutoEncoder (SVAE), which combines Conditional Random Fields with Variational AutoEncoders to address a variety of data modeling challenges. The SVAE has been applied to discrete mixture models, latent linear dynamical systems for video data, and latent switching linear dynamical systems for behavior analysis in video sequences. This approach employs mean field variational inference to approximate the Evidence Lower Bound, targeting specific data types without explicitly focusing on inter-component relationships.

Similarly, Khoshaman and Amin (2018) integrates Boltzmann Machines as priors within VAEs, focusing on discrete variables to model complex and multi-component distributions. Their methodology suggests either factorial or hierarchical structures for the posterior distribution, aiming to effectively model complex and multi-component distributions.

Although significant advances have been made, the application of MRF within the domain of multi-component generative models, particularly in enhancing the integration and modeling of complex dependencies among multiple components remains largely unexplored. Our work seeks to bridge this gap by proposing a novel integration of MRF within a Multi-Component Variational AutoEncoder framework, aimed at capturing the intricate inter-component relationships more effectively. This approach not only leverages the strengths of MRF but also addresses the limitations observed in existing multi-component generative models.

2.4 Diffusion-based Multimodal Generation

More recently, diffusion-based methods have also been explored for multimodal generation, especially in audio–video settings. For instance, Ruan et al. (2023) introduce a joint multimodal diffusion process for synchronized audio–video generation. Xing et al. (2024) connect pretrained unimodal generators through diffusion latent aligners to improve cross-modal generation and alignment. More recently, Hayakawa et al. (2024) propose a discriminator-guided cooperative diffusion

framework for joint audio–video generation.

In this paper, however, we do not further investigate diffusion-based models and instead focus on the VAE family, where our goal is to improve multi-component coherence through explicit structured latent dependencies.

3 METHODS

We define $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ as a collection of M random variables, each representing a distinct component. Our approach employs a Multi-Component Variational AutoEncoder with an integrated Markov Random Field in its latent space, specifically designed to effectively capture the complex inter-component relationships.

3.1 Variational AutoEncoders

Variational AutoEncoders (VAEs) (Kingma and Welling, 2013) use variational inference to approximate intractable posteriors $p(\mathbf{z}|\mathbf{X})$ by maximizing the Evidence Lower Bound (ELBO):

$$\text{ELBO} \;=\; \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}\big[\ln p_\theta(\mathbf{X} \mid \mathbf{z})\big] \;-\; \text{KL}\Big(q_\phi(\mathbf{z} \mid \mathbf{X}) \;\big\|\; p(\mathbf{z})\Big) \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{X})$ is the encoder's variational posterior and $p_\theta(\mathbf{X} \mid \mathbf{z})$ is the decoder's likelihood. In the multi-component setting we adopt a factorized likelihood $p_\theta(\mathbf{X} \mid \mathbf{z}) = \prod_{i=1}^{M} p_{\theta_i}(\mathbf{x}_i \mid \mathbf{z}_i)$. When both $p(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{X})$ are multivariate Gaussians ($\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p$ and $\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q$), the KL divergence admits the closed form

$$\text{KL}(q\|p) = \tfrac{1}{2}\big[\text{tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\delta} - K + \ln\tfrac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|}\big], \tag{2}$$

with $\boldsymbol{\delta} = \boldsymbol{\mu}_p - \boldsymbol{\mu}_q$ and K the total latent dimension. Sampling uses the Cholesky factor $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$: $\mathbf{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so that gradients flow through $(\boldsymbol{\mu}, \mathbf{L})$.

3.2 Markov Random Fields

MRFs offer an intuitive framework to model inter-component dependencies, with nodes representing random variables and edges capturing their relationships.

Mathematically, an MRF is defined over an undirected graph $G = (V, E)$ where each node corresponds to a random variable in the set $(\mathbf{z}) = \{\mathbf{z}_i\}_{i=1}^n$. The joint distribution over these random variables is specified in terms of potential functions over cliques (fully connected subgraphs) of G. A general mathematical definition of an MRF is given by Murphy (2012):

$$p_{MRF} = \frac{1}{\mathcal{Z}} \exp\left[-\sum_{C \in \mathcal{C}} \psi_C(\mathbf{z}_C)\right] \tag{3}$$

where $\mathcal{C}$ is the set of cliques in the graph, ψ_C are the potential functions that map configurations of the random variables within the clique to a real number, $\mathbf{z}_C$ denotes the set of random variables in clique C, and $\mathcal{Z}$ is the partition function that normalizes the distribution. In the context of our work, we model both the prior $p(\mathbf{z})$ and the posterior $q_\phi(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_M)$ as fully connected MRF represented by unary $\psi_i(\mathbf{z}_i)$ and pairwise $\psi_{i,j}(\mathbf{z}_i, \mathbf{z}_j)$ potentials. This leads to the specific form:

$$p_{MRF}(\mathbf{z}) = \frac{1}{\mathcal{Z}} \exp\left[-\left(\sum_{i<j}^{M} \psi_{i,j}(\mathbf{z}_i, \mathbf{z}_j) + \sum_{i}^{M} \psi_i(\mathbf{z}_i)\right)\right] \tag{4}$$

with $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_M)$. This formulation enables the modeling of dependencies between components in our multi-component VAE framework by leveraging the structure of MRFs to capture both local and global interactions within the latent space.

3.3 General MRF MCVAE Architecture and the Gaussian Assumption

3.3.1 General Framework

In order to incorporate the MRF representation into a multi-component VAE framework, we employ the architecture illustrated in Figure 1. Each component $\mathbf{x}_i$ is paired with an encoder-decoder structure: the component-specific encoder outputs the parameters of the unary potentials ψ_i, while a global encoder produces the pairwise potentials $\psi_{i,j}$. Together, these potentials define the joint posterior $q_\phi(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_M)$, capturing both individual component details and their relationships. The latent variable $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$ is sampled from this posterior and decomposed into sub-vectors $\mathbf{z}_i$, which are then reconstructed by their respective decoders. However, the MRF framework introduces two significant challenges: (i) computing the partition function $\mathcal{Z}$ is computationally intractable in fully connected MRFs, and (ii) sampling from generalized MRFs is inherently complex, hindering gradient-based optimization. These limitations motivate our adoption of the Gaussian assumption described next.

3.3.2 GMRF MCVAE

These challenges lead us to adopt a Gaussian assumption for tractability. In standard GMRF notation (Murphy, 2012), the unary and pairwise potentials become $\psi_i(\mathbf{z}_i) = \exp\big(-\frac{1}{2}\,\mathbf{z}_i^\top \Lambda_{i,i}\,\mathbf{z}_i + \eta_i^\top \mathbf{z}_i\big)$ and $\psi_{i,j}(\mathbf{z}_i, \mathbf{z}_j) = \exp\big(-\frac{1}{2}\,\mathbf{z}_i^\top\,\Lambda_{i,j}\,\mathbf{z}_j\big)$, yielding the joint distribution:

$$p_{GMRF}(\mathbf{z}) \;\propto\; \exp\Big(\,\eta^\top\mathbf{z} \;-\; \tfrac{1}{2}\,\mathbf{z}^\top\Lambda\,\mathbf{z}\Big) \tag{5}$$

where $\Lambda = [\Lambda_{i,j}]_{i,j=1}^M$ is the precision matrix and $\eta = (\eta_1, \ldots, \eta_M)$. Defining $\mu = \Lambda^{-1}\eta$ and $\Sigma = \Lambda^{-1}$ yields the standard form $\mathcal{N}(\mu, \Sigma)$, simplifying both understanding and computation. This Gaussian formula-

tion enables differentiable sampling via the reparameterization trick: using Cholesky factorization (Kingma et al., 2019) $\Sigma = L\,L^\top$, we sample as $\mathbf{z} = \mu + L\,\mathbf{u}$ where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, ensuring fully differentiable backpropagation through the latent space.

Prior parameterization. The prior mean $\boldsymbol{\mu}_p$ and covariance $\boldsymbol{\Sigma}_p$ are learnable parameters. We parameterize $\boldsymbol{\Sigma}_p$ through its lower Cholesky factor $\mathbf{L}_p$, with diagonal entries enforced positive via $\mathrm{diag}(\mathbf{L}_p) = \mathrm{softplus}(\tilde{\mathbf{L}}_p)$ on raw parameters $\tilde{\mathbf{L}}_p$, so that $\boldsymbol{\Sigma}_p = \mathbf{L}_p\mathbf{L}_p^\top$ stays symmetric positive definite during training. Encoder-produced covariances use the same Cholesky reparameterization after SPD projection (Algorithm 1). Further implementation notes appear in Appendix A.3.

3.3.3 Conditional Generation

The GMRF assumption also simplifies conditional sampling. In a multivariate Gaussian distribution, conditioning on any subset of variables yields another Gaussian in closed form. Specifically, let $\mathbf{z} = (z_1, \ldots, z_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\Sigma}$ is partitioned into blocks $\Sigma_{i,j}$. Then, for indices $i \neq j$,

$$p(\mathbf{z}_i \,|\, \mathbf{z}_j = z_j) \;=\; \mathcal{N}\big(\hat{\mu}_i,\; \hat{\Sigma}_{i,i}\big) \tag{6}$$

with

$$\hat{\mu}_i = \mu_i + \Sigma_{i,j}\,\Sigma_{j,j}^{-1}(z_j - \mu_j),$$
$$\hat{\Sigma}_{i,i} = \Sigma_{i,i} - \Sigma_{i,j}\,\Sigma_{j,j}^{-1}\,\Sigma_{i,j}^\top$$

A full proof is provided in Appendix A.1. In practice, this closed-form property lets us generate specific components $\mathbf{z}_i$ conditioned on partial observations $\mathbf{z}_j$ without further training or approximation.

3.3.4 Covariance Matrix Construction in the GMRF MCVAE

To incorporate GMRFs into the MCVAE posterior, we parameterize the latent covariance matrix $\Sigma = [\Sigma_{i,j}]_{i,j=1}^{M}$ as a block matrix where $\Sigma_{i,i} \in \mathbb{R}^{d\times d}$ are component-specific variances from individual encoders, and $\Sigma_{i,j}$ $(i \neq j)$ are cross-component covariances from the global encoder. We ensure Σ is symmetric positive definite (SPD) using the following theorem:

Theorem 1. *Consider a block matrix Σ as defined in Section 3.3.4. If for each $i \in \{1, \ldots, M\}$ $\Sigma_{i,i}$ is SPD and satisfies:*

$$||\Sigma_{i,i}^{-1}||^{-1} \geq \sum_{k\neq i} ||\Sigma_{i,k}||$$

where $||.||$ is the spectral norm, then Σ is also SPD.

Algorithm 1 SPD Covariance Matrix Construction

Require: $\{\Sigma_{i,i}\}_{i=1}^{M}$ (SPD), $\{\tilde{\Sigma}_{i,j}\}_{i\neq j}$, $\delta > 0$, $\epsilon < 1$
Ensure: SPD matrix Σ
for $i = 1$ to M **do**
  $s_i \leftarrow \sum_{j\neq i} ||\tilde{\Sigma}_{i,j}|| + \delta$
  $\alpha_i \leftarrow \min(1, \epsilon||\Sigma_{i,i}^{-1}||^{-1}/s_i)$
end for
for all $i, j \in \{1, \ldots, M\}$ **do**
  $\Sigma_{i,j} \leftarrow \begin{cases} \Sigma_{i,i} & \text{if } i = j \\ \tilde{\Sigma}_{i,j}\sqrt{\alpha_i\alpha_j} & \text{otherwise} \end{cases}$
end for

The complete proof is provided in Appendix A.2. This result establishes the theoretical foundation for our SPD matrix construction method, which we present in Algorithm 1. We defer the algorithmic details and computational complexity analysis to Appendix A.5.

4 EXPERIMENTS

We benchmark the proposed GMRF MCVAE against four leading multi-component Variational AutoEncoders: MVAE (Wu and Goodman, 2018), MMVAE (Shi et al., 2019), MoPoE-VAE (Sutter et al., 2021), and MMVAE+ (Palumbo et al., 2023). The core objective is to evaluate the models' ability to learn and preserve complex inter-component structure.

Datasets. We evaluate our GMRF MCVAE on three diverse benchmarks that test different aspects of multi-component modeling. The **Copula dataset** is a synthetic benchmark we designed using Gaussian Copulas to create components with intricate statistical dependencies while maintaining simple individual distributions $\mathcal{U}([0, 1])$, isolating the challenge of dependency modeling. **PolyMNIST** (Sutter et al., 2021) is an established benchmark where five modalities show the same digit with different styles and backgrounds, testing component complexity with simple inter-dependencies. **BIKED** (Regenwetter et al., 2022) contains real-world bicycle designs with five essential components, requiring models to maintain structural and spatial coherence for industrial feasibility. These datasets span from controlled synthetic dependencies to complex real-world structural constraints.

We average all numerical results over three independently trained models. Key tables report mean ± standard deviation across runs. Training details, including architectures and hyperparameters, are provided in the Appendix A.3. Wall-clock training time per epoch on PolyMNIST is reported in Appendix A.6: despite the covariance construction overhead, our model

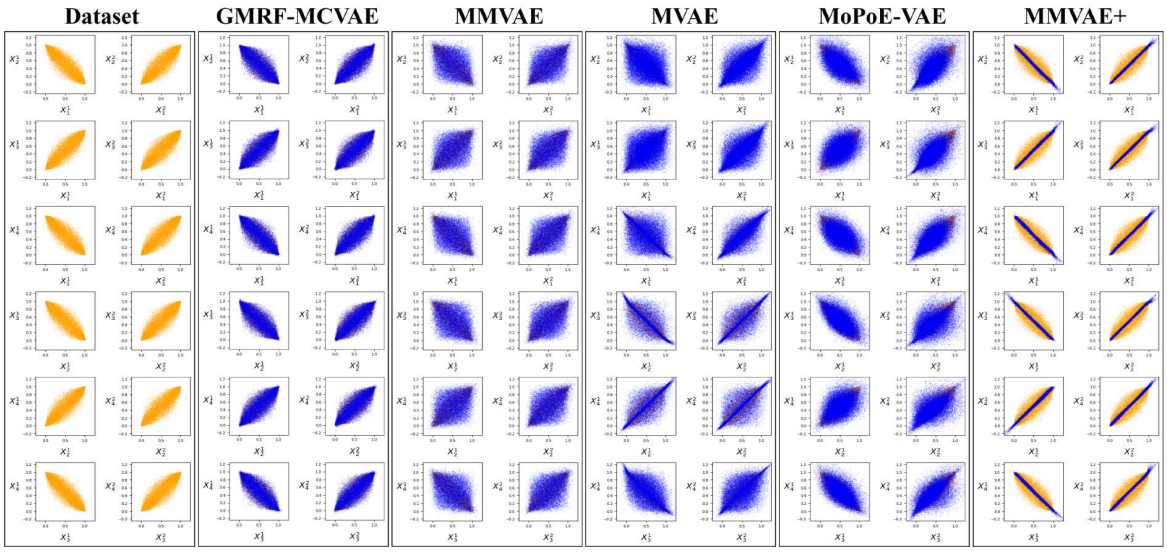

Figure 2: Qualitative results for the unconditional generations on the Copula dataset. Each subplot visualizes joint distributions for each pair of coordinates $(\mathbf{x}_i^1, \mathbf{x}_j^1)$ and $(\mathbf{x}_i^2, \mathbf{x}_j^2)$ across the four two-dimensional components $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$. The true distributions are depicted in orange and the generated ones in blue.

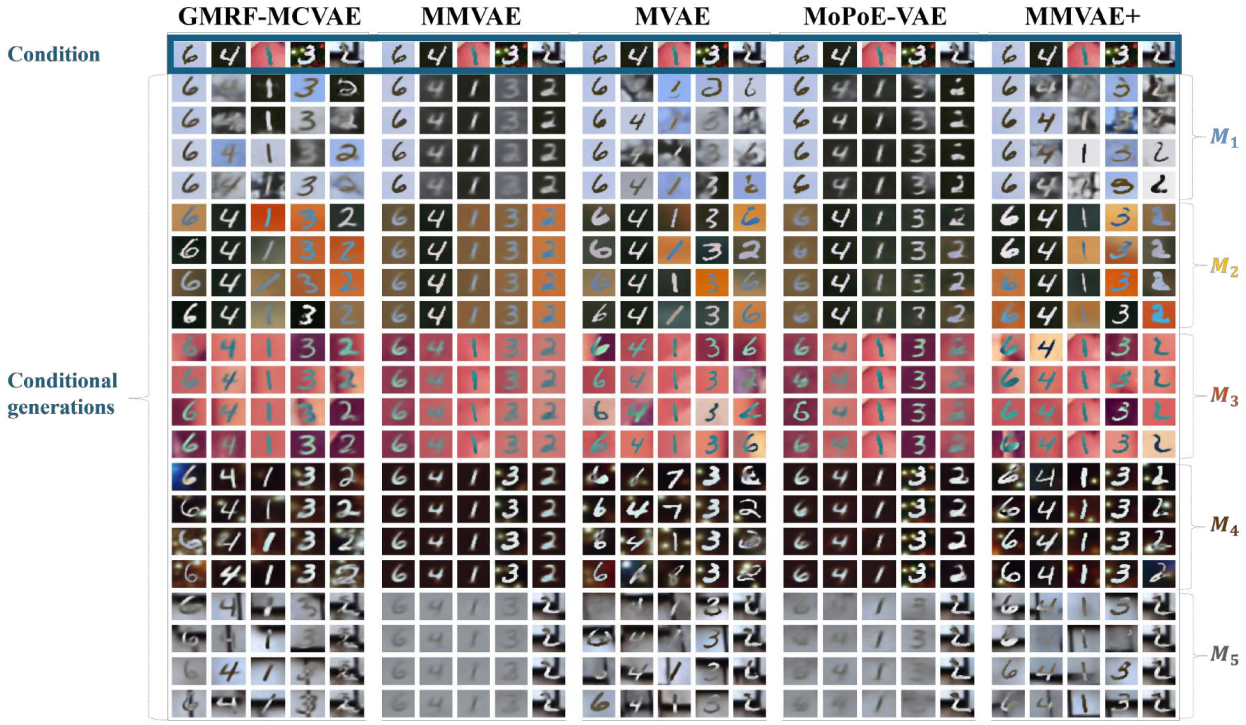

Figure 3: PolyMNIST conditional generations. Each block corresponds to a model. In each column, the first image corresponds to the condition, followed by the conditionally generated components M_i.

is faster per epoch than the other baselines.

4.1 Copula Dataset Evaluation

This subsection evaluates the capability of each model to handle and represent complex inter-component interactions.

The synthetic dataset consists of four two-dimensional components, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, each defined as $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2)$ where each component $\mathbf{x}_i^j$ is uniformly distributed, $\mathbf{x}_i^j \sim \mathcal{U}([0,1])$, for $i \in \{1,2,3,4\}$ and $j \in$

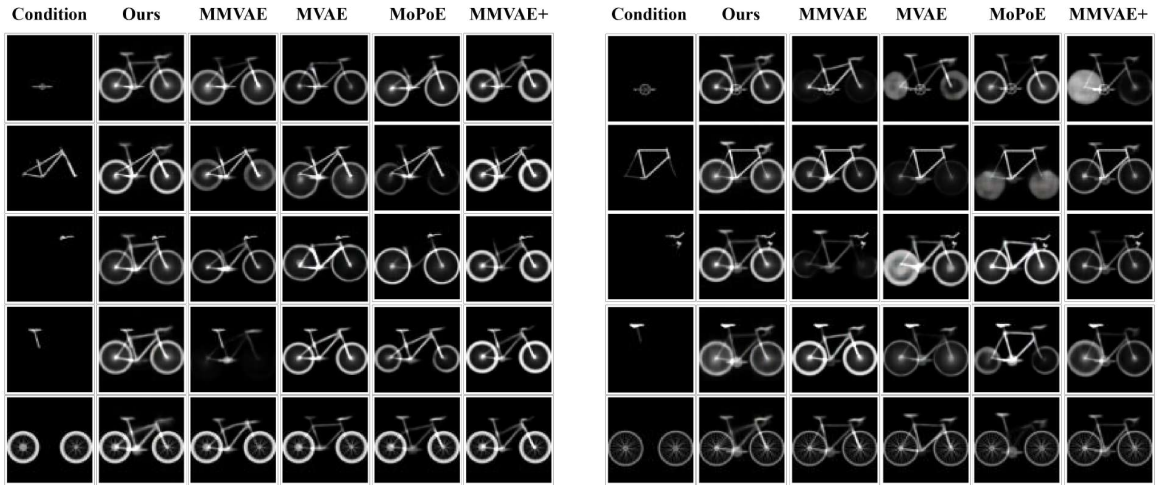

Figure 4: Conditional generation on BIKED. The first column shows the conditioning components, while each subsequent column presents the remaining generated components for each model, overlaid to form the complete bike.

Model	Uncond. Gen.			Cond. Gen.		
	Dim1	Dim2	Mean	Dim1	Dim2	Mean
MVAE	2.7±0.4	3.2±0.8	2.9±0.5	3.0±0.3	3.1±0.7	3.1±0.4
MMVAE	5.2±0.5	4.5±0.4	4.8±0.6	5.4±0.1	4.7±0.3	5.0±0.04
MoPoE	1.9±0.5	2.6±0.6	2.2±0.4	6.0±0.7	5.6±0.1	5.9±0.3
MMVAE+	8.1±0.05	4.9±0.07	6.5±0.06	5.2±0.2	4.9±0.1	5.1±0.1
Ours	**0.7**±0.1	**0.95**±0.01	**0.86**±0.04	**2.6**±0.2	**2.7**±0.1	**2.6**±0.1

Table 1: Mean ± std over 3 runs. Wasserstein distance ($\times 10^3$) on the synthetic Copula dataset for unconditional and conditional generation.

$\{1, 2\}$. The coordinates of each component are generated using two Gaussian Copulas, $C_j(\mathbf{x}_1^j, \ldots, \mathbf{x}_4^j)$, with uniform means $\mu_j = [3, \ldots, 3]$ and standard deviations $\sigma_j = [1, \ldots, 1]$. The correlation matrices R^j have off-diagonal elements set as $R^j_{k,l} = ((-1)^j)^{k+l} \cdot 0.9$ (Figure 2).

Metric We assess model performance using the Wasserstein distance (Villani, 2009), which measures the optimal transport cost between the empirical probability density functions (PDFs) of the generated and true samples for each component's coordinates. This metric captures differences in both the supports and shapes of distributions. The average of these distances across all comparisons serves as an aggregate performance measure.

4.1.1 Qualitative Comparison

In the analysis of joint distributions (Figure 2), the GMRF MCVAE demonstrates superior alignment with the true distribution, indicating high-quality generations. The MVAE captures certain inter-component relationships more effectively, although this varies across components. Notably, the third component appears less accurate. This variability is explored in further detail in appendix A.9. As noted by Shi et al. (2019), this may stem from the "veto" effect, where experts with higher precision dominate the joint posterior, leading to biased predictions. The MoPoE-VAE combines features of both the MVAE and MMVAE, producing less noisy images but failing to consistently enforce cross-component relationships. Notably, MMVAE+ underperforms compared to other models. While it captures the general correlation between components, it struggles to model the complete dependency structure of the true distribution. This may be due to the increased complexity introduced by component-specific sampling, which complicates the representation of intricate inter-component dependencies.

4.1.2 Quantitative Comparison

Table 1 confirms the observations from the previous section. The Wasserstein distances between the true and generated distribution PDFs indicate that the GMRF MCVAE generates distributions closer to the true distributions.

4.2 PolyMNIST Dataset Evaluation

In this section, we present the results of our evaluation on the PolyMNIST dataset, assessing the models' ability to handle component complexity and ensure consistency across components. PolyMNIST (Sutter et al., 2021) extends MNIST into a five-component dataset, where each data point consists of five label-consistent images of the same digit, rendered in distinct handwriting styles and paired with unique background types. This setup challenges models to capture

Model	Unconditional		Conditional				
	Comp. FID ↓	Full FID ↓	Comp. FID ↓	Full FID ↓	Comp. SSIM ↑	Full SSIM ↑	WS ↓
MVAE	136.60±1.82	155.06±2.13	134.78±0.81	151.77±3.14	0.85±0.02	0.53±0.02	0.46±0.03
MMVAE	**133.59±0.78**	**148.34±1.68**	132.45±0.69	**146.92±1.05**	0.84±0.01	0.50±0.01	0.47±0.01
MoPoE	136.29±3.05	158.25±2.42	131.94±1.77	150.44±1.13	0.85±0.01	0.53±0.01	0.49±0.02
MMVAE+	137.17±1.74	176.12±1.32	131.78±1.21	171.48±2.29	0.84±0.02	0.54±0.03	0.51±0.09
Ours	136.83±0.93	186.88±1.40	**131.52±0.83**	173.70±1.30	**0.88±0.02**	**0.64±0.02**	**0.31±0.01**

Table 2: Mean ± std (3 runs). FID, SSIM, and center-of-mass Wasserstein (WS) on BIKED. Comp./Full denote component-wise and full-overlap images.

the shared digit identity while managing the complexity introduced by handwriting and background variations. As detailed in Appendix A.3.1, we mask 75% of off-diagonal covariance parameters so the number of latent parameters remains comparable to the baselines.

Metrics We evaluate model performance using the Fréchet Inception Distance (FID) (Heusel et al., 2017), which measures generative quality and diversity (Ho and Salimans, 2022), and the Structural Similarity Index (SSIM) (Wang et al., 2004), which assesses perceptual similarity between generated and real images. Additionally, we evaluate global coherence using the methodology proposed by Palumbo et al. (2023), which measures whether all generated components share the same digit.

Conditional FID protocol. Following MMVAE (Shi et al., 2019) and MMVAE+ (Palumbo et al., 2023), we select one component as observed, encode it, and generate the remaining components. Observed and generated parts are overlaid into a single composite image; FID is then computed between these composites and real test images. All baselines follow the same procedure. SSIM and coherence use the same composite representation.

Qualitative Comparison We observe that The GMRF MCVAE model produces consistently complete digits, though slightly blurrier than the MMVAE+ and MVAE (cf Figure 3). The MVAE generates sharp images but struggles to align digit identities across components, likely due to the conditional independence assumption induced by the PoE aggregation. MMVAE however suffers from an averaging effect, which dilutes component-specific details and leads blend backgrounds and a lack of digit variability. MoPoE-VAE by construction attempts to balance PoE and MoE, but visually exhibits their combined drawbacks: oversmoothing and incomplete digits. Despite some tradeoffs in sharpness, our GMRF MCVAE consistently preserves the coherence and diversity of the digits, ensuring high generative quality and semantic coherence.

Model	Unconditional		Conditional		
	FID ↓	Coh. ↑	FID ↓	Coh. ↑	SSIM ↑
MVAE	95.14±0.40	0.139±0.010	94.71±3.20	0.448±0.024	0.993±0.001
MMVAE	170.87±3.17	0.175±0.010	198.80±2.69	0.517±0.010	0.995±0.001
MoPoE-VAE	106.12±1.96	0.018±0.001	162.74±4.12	0.475±0.006	0.995±0.001
MMVAE+	**87.23±1.41**	0.210±0.006	**82.05±0.78**	0.856±0.013	0.994±0.001
Ours	118.21±1.71	**0.321±0.014**	180.76±3.11	**0.869±0.016**	**0.995±0.001**

Table 3: Mean ± std (3 runs). Results on PolyMNIST: FID, coherence (Coh.), and SSIM.

Quantitative Comparison The quantitative evaluation, as presented in Table 3, reveals that the GMRF MCVAE model performs competitively across all considered metrics on the PolyMNIST dataset. Although the MVAE model achieves the lowest FID scores, the GMRF MCVAE exhibits higher values of cross-coherence and SSIM, suggesting enhanced preservation of structural integrity and global coherence in the generated samples. These results underscore the GMRF MCVAE's ability to produce high-quality, structurally coherent outputs, indicating its robustness in multi-component generative modeling.

4.3 BIKED Evaluation

BIKED (Regenwetter et al., 2022) dataset is a collection of 4,500 bicycle models designed by hundreds of designers, originally introduced for data-driven bicycle design applications. The dataset includes various types of design information, such as parametric data, class labels, and images of full bike assemblies and individual components. In this work, we focus on the segmented component images, where each image corresponds to a different part of a bicycle. Specifically, we retain only the five essential components (saddle, frame, crank, wheels, and handlebars), while discarding components that do not consistently appear across all images, such as water bottles and cargo racks. To reduce GPU load and accelerate training, we convert all images to grayscale and downsample them to 64×64 resolution.

Metrics We evaluated generation quality and diversity using FID and structural coherence with SSIM, which measures how well the overall structure of conditionally generated bikes aligns with real ones.

While SSIM provides a perceptual evaluation of global shape consistency, industrial design often requires more domain-specific geometric constraints to ensure practical feasibility, e.g., maintaining an appropriate crank-to-wheel distance for maneuverability. To capture such structural dependencies in a scalable and generalizable manner, we propose the Wasserstein distance between component centers of mass, which quantifies the global spatial distribution of parts. Concretely, for each sample we compute the 2D center of mass of every component, stack them into $\mathbf{c} \in \mathbb{R}^{2M}$, and report the 1-Wasserstein distance between empirical distributions of $\mathbf{c}$ for generated and real images. This approach enables us to assess whether the generated components form structurally plausible assemblies, independent of predefined expert rules, making it applicable to other industrial multi-component datasets.

Conditional FID on BIKED. We condition on each of the five components in turn, generate the remaining four, and overlay all five into a full-bike image. FID is averaged over the five conditioning choices. SSIM and the center-of-mass Wasserstein metric use the same overlay representation, with all baselines evaluated identically.

4.3.1 Qualitative comparison

Qualitative results of conditional generation (Figure 4) highlight the limitations of the baseline models. Although MVAE, MMVAE, and MoPoE-VAE exhibit high variability by mixing components from different types of bikes, their generated outputs often lack structural integrity. Specifically, the outputs frequently contain missing or incomplete components, and MVAE, in particular, produces structurally infeasible designs, such as misaligned rear frame sections, causing the rear wheel to be improperly positioned, making the generated bikes impractical for real-world assembly. MMVAE+ demonstrates better structural consistency, as its generations more closely resemble complete and coherent bicycles when conditioned on specific components. However, this comes at the cost of reduced variability, and we observe frequent missing components, such as absent saddles. Our model, by contrast, achieves a balance between variability and structural integrity. While its generations may not be as diverse as those of MVAE, they remain structurally coherent and plausible as functional bicycle designs, aligning with expected industrial feasibility.

4.3.2 Quantitative comparison

The results in Table 2 align with our qualitative observations. MVAE, MMVAE, and MoPoE-VAE achieve the lowest FID scores, indicating higher diversity in their generations. However, our model maintains competitive FID performance while achieving the best SSIM and Wasserstein distance scores, confirming its ability to generate structurally coherent designs. Higher SSIM values indicate that our model better preserves local structural consistency in conditional generations, ensuring that the generated components align more closely with real designs. Additionally, the significantly lower Wasserstein distance between the center of mass of the components suggests superior spatial integrity, meaning that the generated parts are correctly positioned relative to each other. This supports our model's ability to balance diversity with structural plausibility, producing generations that are not only varied but also geometrically coherent and physically plausible.

Block-averaged posterior correlations reveal that mechanically coupled pairs consistently exhibit higher correlation than loosely related ones; see Appendix A.8 for details.

5 CONCLUSION

This work addressed a core limitation of current multi-component generative models: their inability to explicitly capture dependencies between components. We introduced GMRF MCVAE, integrating Gaussian Markov Random Fields within both prior and posterior distributions to enable structured latent spaces that directly model cross-component relationships.

Our model achieved competitive results on PolyMNIST and outperformed state-of-the-art baselines on both synthetic Copula and real-world BIKED datasets, demonstrating its strength where structural coherence is essential. GMRF MCVAE generates assemblies preserving both diversity and spatial integrity, crucial for industrial design applications.

Future work will investigate sparser MRF structures, improved interpretability, and extensions to diffusion-based generative models.

Acknowledgements

This work was carried out at Centre Borelli, ENS Paris-Saclay, Université Paris-Saclay, CNRS, in collaboration with Manufacture Française des Pneumatiques Michelin. We thank the anonymous reviewers for their constructive feedback.

References

Bach, F. and Jordan, M. (2002). Learning graphical models with mercer kernels. *Advances in Neural Information Processing Systems*, 15.

Bello, M. G. (1994). A combined markov random field and wave-packet transform-based approach for image segmentation. *IEEE transactions on image processing*, 3(6):834–846.

Carreira-Perpinan, M. A. and Hinton, G. (2005). On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR.

Cobb, A., Roy, A., Elenius, D., Heim, F., Swenson, B., Whittington, S., Walker, J., Bapty, T., Hite, J., Ramani, K., et al. (2023). Aircraftverse: a large-scale multimodal dataset of aerial vehicle designs. *Advances in Neural Information Processing Systems*, 36:44524–44543.

Feingold, D. G. and Varga, R. S. (1962). Block diagonally dominant matrices and generalizations of the gerschgorin circle theorem.

Hayakawa, A., Ishii, M., Shibuya, T., and Mitsufuji, Y. (2024). Mmdisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation. *arXiv preprint arXiv:2405.17842*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29.

Khoshaman, A. H. and Amin, M. (2018). Gumbolt: Extending gumbel trick to boltzmann priors. *Advances in Neural Information Processing Systems*, 31.

Kindermann, R. and Snell, J. L. (1980). *Markov random fields and their applications*, volume 1. American Mathematical Society.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Komodakis, N. and Tziritas, G. (2007). Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Transactions on Image Processing*, 16(11):2649–2661.

Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.

Lee, S. I. and Yoo, S. J. (2020). Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing*, 76:8294–8312.

Mbacke, S. D., Clerc, F., and Germain, P. (2024). Statistical guarantees for variational autoencoders using pac-bayesian theory. *Advances in Neural Information Processing Systems*, 36.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Oubari, F., Meunier, R., Décatoire, R., and Mougeot, M. (2024). A meta-vae for multi-component industrial systems generation. In Arai, K., editor, *Intelligent Computing*, pages 234–251, Cham. Springer Nature Switzerland.

Palumbo, E., Daunhawer, I., and Vogt, J. E. (2023). Mmvae+: Enhancing the generative quality of multimodal vaes without compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview.

Perez, P. et al. (1998). *Markov random fields and images*, volume 469. IRISA.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Puhr-Westerheide, D., Cyran, C. C., Sargsyan-Bergmann, J., Todica, A., Gildehaus, F.-J., Kunz, W. G., Stahl, R., Spitzweg, C., Ricke, J., and Kazmierczak, P. M. (2019). The added diagnostic value of complementary gadoxetic acid-enhanced mri to 18 f-dopa-pet/ct for liver staging in medullary thyroid carcinoma. *Cancer Imaging*, 19:1–10.

Rahimi, A., Khalil, A., Faisal, A., and Lai, K. W. (2022). Ct-mri dual information registration for the diagnosis of liver cancer: A pilot study using point-based registration. *Current medical imaging*, 18(1):61–66.

Regenwetter, L., Curry, B., and Ahmed, F. (2022). Biked: A dataset for computational bicycle design with machine learning benchmarks. *Journal of Mechanical Design*, 144(3):031706.

Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N. J., Jin, Q., and Guo, B. (2023). Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10219–10228.

Shi, Y., Paige, B., Torr, P., et al. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Sutter, T., Daunhawer, I., and Vogt, J. (2020). Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110.

Sutter, T. M., Daunhawer, I., and Vogt, J. E. (2021). Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M., and Witten, D. (2014). Learning graphical models with hubs. *arXiv preprint arXiv:1402.7349*.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. (2017). Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Villani, C. (2009). The wasserstein distances. *Optimal Transport: Old and New*, pages 93–111.

Vuffray, M., Misra, S., and Lokhov, A. (2020). Efficient learning of discrete graphical models. *Advances in Neural Information Processing Systems*, 33:13575–13585.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Welling, M. and Sutton, C. (2005). Learning in markov random fields with contrastive free energies. In *International Workshop on Artificial Intelligence and Statistics*, pages 397–404. PMLR.

Wu, M. and Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., and Huang, J. (2024). Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36.

Xing, Y., He, Y., Tian, Z., Wang, X., and Chen, Q. (2024). Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] NVIDIA RTX 3060 Laptop GPU (6 GB); wall-clock timings in Appendix A.6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [No]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A SUPPLEMENTARY MATERIAL

A.1 Conditional Sampling

The conditional distribution of a normally distributed random variable given another is also normally distributed. This is known for the bivariate case in the Matrix Cookbook (Petersen et al., 2008). We extend this result for $n \geq 2$, considering a random vector $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the following probability density function:

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix} \right). \tag{7}$$

For any pair of indices $i \neq j$ from the set $\{1, \ldots, n\}$, the conditional distribution of $\mathbf{z}_i$ given z_j is

$$p(\mathbf{z}_i | \mathbf{z}_j = z_j) = \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_{i,i}), \tag{8}$$

where

$$\begin{cases} \hat{\mu}_i = \mu_i + \Sigma_{i,j}\Sigma_{j,j}^{-1}(z_j - \mu_j), \\ \hat{\Sigma}_{i,i} = \Sigma_{i,i} - \Sigma_{i,j}\Sigma_{j,j}^{-1}\Sigma_{j,i}. \end{cases} \tag{9}$$

Proof. Consider a random vector $\mathbf{z}$ and distinct indices i and j. Define the transformation $\mathbf{y} = A\mathbf{z}_i + B\mathbf{z}_j$ such that $\mathbf{y}$ and $\mathbf{z}_j$ are independent. To achieve $\text{cov}(\mathbf{y}, \mathbf{z}_i) = 0$, it follows that

$$A\Sigma_{i,j} + B\Sigma_{j,j} = 0.$$

Selecting $A = I$, leads to

$$B = -\Sigma_{i,j}\Sigma_{j,j}^{-1}.$$

Substituting back, we obtain

$$\mathbf{y} = \mathbf{z}_i - \Sigma_{i,j}\Sigma_{j,j}^{-1}\mathbf{z}_j.$$

The independence implies $\mathbf{E}[\mathbf{y}|\mathbf{z}_j] = \mathbf{E}[\mathbf{y}] = \mu_i$. Consequently, the conditional expectation of $\mathbf{z}_i$ given z_j is

$$\begin{aligned} \mathbf{E}[\mathbf{z}_i | z_j] &= \mathbf{E}[\mathbf{y} + \Sigma_{i,j}\Sigma_{j,j}^{-1}\mathbf{z}_j | z_j] \\ &= \mathbf{E}[\mathbf{y}|z_j] + \Sigma_{i,j}\Sigma_{j,j}^{-1} z_j \\ &= \mu_i + A(\mu_j - z_j). \end{aligned}$$

For the variance, we derive:

$$\begin{aligned} \text{var}(\mathbf{z}_i | z_j) &= \text{var}(\mathbf{y} - B\mathbf{z}_j | z_j) \\ &= \text{var}(\mathbf{y}|z_j) + \text{var}(B\mathbf{z}_j | z_j) \\ &\quad - B\text{cov}(\mathbf{y}, -\mathbf{z}_j) - \text{cov}(\mathbf{y}, -\mathbf{z}_j)B' \\ &= \text{var}(\mathbf{y}|z_j) \\ &= \text{var}(\mathbf{y}). \end{aligned}$$

Thus:

$$\begin{aligned} \text{var}(\mathbf{z}_i | z_j) &= \text{var}(\mathbf{z}_i + B\mathbf{z}_j) \\ &= \text{var}(\mathbf{z}_i) + B\text{var}(\mathbf{z}_j)B' + B\text{cov}(\mathbf{z}_j, \mathbf{z}_i) \\ &\quad - \text{cov}(\mathbf{z}_i, \mathbf{z}_j)B' \\ &= \Sigma_{i,i} + B\Sigma_{j,j}B' - B\Sigma_{j,i} - \Sigma_{i,j}B' \\ &= \Sigma_{i,i} - \Sigma_{i,j}\Sigma_{j,j}^{-1}\Sigma_{j,i} \end{aligned}$$

This final expression for $\text{var}(\mathbf{z}_i | \mathbf{z}_j = z_j)$ is the variance of the conditional distribution $p(\mathbf{z}_i | \mathbf{z}_j = z_j) = \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_{i,i})$, where $\hat{\Sigma}_{i,i} = \Sigma_{i,i} - \Sigma_{i,j}\Sigma_{j,j}^{-1}\Sigma_{j,i}$. □

A.2 Demonstration of Theorem 1

A.2.1 Preliminaries

This section demonstrates how our architecture can generate full diagonal block covariance matrices while ensuring symmetric positive definiteness.

Let M be a complex matrix, and let E and F be two vector spaces equipped with norms $\|\cdot\|_E$ and $\|\cdot\|_F$, respectively, such that for all $\mathbf{x} \in E$, $M\mathbf{x} \in F$.

The standard operator norm of M is defined as:

$$\|M\| = \sup_{\mathbf{x}\neq 0} \frac{\|M\mathbf{x}\|_F}{\|\mathbf{x}\|_E}.$$

Consider a block matrix Σ defined as:

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix}, \tag{10}$$

where $\Sigma_{i,i}$ are square matrices and $\Sigma_{i,j}$ for $i \neq j$ are rectangular matrices, all with complex entries. In the remainder of this section, we will refer to Σ as defined in Equation 10.

Following Feingold and Varga (1962), we say that Σ is block diagonally dominant with respect to the norm $||\cdot||$ if:

$$||\Sigma_{i,i}^{-1}||^{-1} \geq \sum_{\substack{k=1 \\ k\neq i}}^{n} ||\Sigma_{i,k}||, \quad \forall i \in \{1,\ldots,n\}. \tag{11}$$

In order to proove Theorem 1, we need to present first the following two key theorems from (Feingold and Varga, 1962):

Theorem 2. *A block matrix Σ is nonsingular if it is block strictly diagonally dominant (i.e., strict inequality holds in Equation 11).*

Theorem 3. *For a block matrix Σ, each eigenvalue λ satisfies:*

$$\left(||(\Sigma_{i,i} - \lambda\mathbf{I})^{-1}||\right)^{-1} \leq \sum_{\substack{k=1 \\ k\neq i}}^{n} ||\Sigma_{i,k}||,$$

for at least one $i \in \{1,\ldots,n\}$.

A.2.2 Proof of Theorem 1

We outline the logic before the detailed steps: block strict diagonal dominance (Theorem 2) gives nonsingularity; Theorem 3 locates each eigenvalue λ near some diagonal block spectrum; combining the spectral-norm bounds shows every eigenvalue must be positive because each diagonal block is SPD. The matrix is therefore SPD.

Proof. To prove Theorem 1, we show that Σ is positive definite.

1. *Nonsingularity*: From Theorem 2, the block strictly diagonal dominance of Σ guarantees that it is nonsingular.

2. *Positivity*: Let λ be an eigenvalue of Σ. Theorem 3 ensures that there exists at least one $i \in \{1,\ldots,n\}$ such that:

$$\left(||(\Sigma_{i,i} - \lambda\mathbf{I})^{-1}||\right)^{-1} \leq \sum_{\substack{k=1 \\ k\neq i}}^{n} ||\Sigma_{i,k}||.$$

Since we consider in our theorem that $||\cdot||$ is the spectral norm, we have:

$$||(\Sigma_{i,i} - \lambda\mathbf{I})^{-1}|| = \sup_{j\in\{1,\ldots,d\}} \left|\frac{1}{\sigma_j^i - \lambda}\right|,$$

where $(\sigma_j^i)_j$ are the eigenvalues of $\Sigma_{i,i}$. Let $k \in \{1, \dots, d\}$ be the index where the supremum is achieved. Substituting this into the inequality gives:

$$\left(||(\Sigma_{i,i} - \lambda \mathbf{I})^{-1}||\right)^{-1} = |\sigma_k^i - \lambda|.$$

Using Theorem 2, we have:

$$|\sigma_k^i - \lambda| \leq \sum_{\substack{k=1 \\ k \neq i}}^{n} ||\Sigma_{i,k}|| < ||\Sigma_{i,i}^{-1}||^{-1}.$$

For the spectral norm:

$$||\Sigma_{i,i}^{-1}|| = \frac{1}{\sigma_{\min}^i},$$

where $\sigma_{\min}^i$ is the smallest eigenvalue of $\Sigma_{i,i}$. Substituting into the inequality, we get:

$$|\sigma_k^i - \lambda| < \sigma_{\min}^i.$$

This implies:

$$-\sigma_{\min}^i + \sigma_k^i < \lambda < \sigma_{\min}^i + \sigma_k^i.$$

Since $-\sigma_{\min}^i + \sigma_k^i \geq 0$, it follows that $\lambda > 0$, proving that all eigenvalues of Σ are positive. And thus, Σ is SPD, completing the proof.

□

A.3 Technical details for the experiments

Throughout all the experiments, we train each model on 3 independent initializations. In this section we provide the experimental details for both PolyMNIST and the Copula experiments.

A.3.1 PolyMNIST & BIKED Experiments

We employ consistent encoder/decoder architectures across all baseline models, using publicly available implementations for MVAE, MMVAE, and MoPoE-VAE from (Sutter et al., 2020), and for MMVAE+ from (Palumbo et al., 2023). We use similar encoders/decoders architectures in both experiments with different parameters reported on table 8. Our GMRF MCVAE follows a similar ResNet-based design but differs in that we do not employ two separate latent spaces for joint and component specific encodings. Instead, we introduce an additional fully connected network (three layers of 128 ReLU units each, followed by a linear output) to generate the off-diagonal covariance blocks.

To ensure fair comparisons in both the PolyMNIST and BIKED tasks, we configure all factorized baseline models with a 32-dimensional latent space for PolyMNIST and an 8-dimensional latent space for BIKED (both split into shared and modality-specific subspaces). For unfactorized baselines, we use latent spaces of 512 dimensions for PolyMNIST and 16 dimensions for BIKED. In contrast, our GMRF MCVAE architecture employs a fixed latent space dimension of 16 for PolyMNIST and 4 for BIKED. Additionally, in the PolyMNIST experiment, we mask out 75% of the off-diagonal parameters in the covariance matrix of the GMRF MCVAE, resulting in 660 parameters per distribution, to align with the effective latent capacity of the baselines.

All models are trained for 100 epochs and monitored using coherence and Fréchet Inception Distance (FID). For baseline models, we experiment with $\beta \in \{5, 2.5, 1\}$ (lower/higher beta values result in poorer performances) . For the GMRF MCVAE in particular, we explore $\beta \in \{2.5 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$, where β controls the KL term weight in the ELBO (Higgins et al., 2017). We find that $\beta = 1 \times 10^{-3}$ yields the best performance in both datasets.

A.3.2 Copula Dataset Experiment

Table 4 presents the architecture details for both the encoders and decoders used in the Copula experiment. To maintain consistency in latent capacities across different models, the GMRF MCVAE was configured with a latent dimension of 2 (yielding a total capacity of 44), while all other models used a latent dimension of 3 (total

capacity of 48). All models were trained for 200 epochs, exploring a range of β values: $\{2.5, 1, 0.1, 0.05, 0.001\}$. For baseline models, both Gaussian and Laplacian distributions were tested for the prior, posterior, and log-likelihood calculations. Factorized and unfactorized variants were evaluated for MVAE, MMVAE, and MoPoE-VAE.

Component	Layer	Units	Activation
Encoder	Fully Connected	2×256	ReLU
	Fully Connected	256×256	ReLU
	Fully Connected - σ_{shared}	$256 \times latent$	Linear
	Fully Connected - $logvar_{shared}$	$256 \times latent$	Linear
	(if factorized) Fully Connected - $\sigma_{specific}$	$256 \times latent$	Linear
	(if factorized) Fully Connected - $logvar_{specific}$	$256 \times latent$	Linear
Decoder	Fully Connected	$input \times 256$	ReLU
	Fully Connected	256×256	ReLU
	Fully Connected	256×2	Linear

Table 4: Architecture details of the encoders and decoders used in the Copula experiment.

Layer	Input Channels	Output Channels	Kernel Size	Stride	Note
Conv2d (conv_0)	fin	$fhidden$	3×3	1	Padding=1
Activation	–	–	–	–	LeakyReLU (0.2)
Conv2d (conv_1)	$fhidden$	$fout$	3×3	1	Padding=1, bias enabled
Activation	–	–	–	–	LeakyReLU (0.2)
Shortcut (if $fin \neq fout$)	fin	$fout$	1×1	1	Learned (no padding)

Table 5: ResNet Block Architecture. Each block consists of two convolutional layers with LeakyReLU activations and an optional learned shortcut when the input and output channel dimensions differ.

Stage	Layer Type	Output Dimensions	Activation
Input	Image	Input size	–
Initial Conv	Conv2d ($3 \to 32$, kernel=3, padding=1)	Input size	–
ResNet Blocks	Sequence of ResNet blocks with AvgPool (3 blocks)	$nf_0 \times s_0 \times s_0$	LeakyReLU (0.2)
Flatten	View into vector	$nf_0 \cdot s_0^2$	–
FC Layers mod	Two separate fully-connected layers	Latent Dimension	–
FC Layers joint	Two separate fully-connected layers	Latent Dimension	–
FC Layer off-diag	One fully-connected layer	Latent Dimension	–

Table 6: Encoder Architecture for PolyMNIST and BIKED. The encoder first applies a convolution to the input image, followed by a series of ResNet blocks with average pooling. The final feature map is flattened and processed by fully-connected layers to generate the latent mean and diagonal covariance parameters. For all baseline models except ours, the "FC Layers mod" and "FC Layers joint" correspond to the modality-specific and shared posterior encodings, respectively. In contrast, our architecture uses an additional fully connected layer "FC Layer off-diag" to encode the embedding utilized by the global encoder for generating the off-diagonal elements of the covariance matrix.

A.4 Diagonal bloc construction details

We remind that the proposed SDP construction algorithm:

1. **Diagonal Blocks**: Each component-specific encoder produces and SPD $\Sigma_{i,i}$.

2. **Off-Diagonal Blocks**: The global encoder outputs the off-diagonal blocks $\tilde{\Sigma}_{i,j} \in \mathbb{R}^{d \times d}$, capturing relationships between components. By symmetry, $\tilde{\Sigma}_{j,i} = \tilde{\Sigma}_{i,j}^\top$. Stacking these blocks yields:

$$\tilde{\Sigma} = \begin{bmatrix} 0 & \cdots & \tilde{\Sigma}_{1,M} \\ \vdots & \ddots & \vdots \\ \tilde{\Sigma}_{M,1} & \cdots & 0 \end{bmatrix} \tag{12}$$

Stage	Layer Type	Output Dimensions	Activation
Input	Latent vector	Latent Dimension (*2 for all models but ours)	–
FC Layer	Fully-connected	$nf_0 \times s_0 \times s_0$	–
Reshape	Reshape to tensor	$256 \times 8 \times 8$	–
Upsampling	Sequential ResNet blocks with Upsample (scale factor=2)	Gradually upsample to $nf \times 64 \times 64$	LeakyReLU (0.2)
Output Conv	Conv2d (from nf to 3 channels, kernel=3, padding=1)	Output size	–

Table 7: Decoder Architecture for PolyMNIST. The decoder maps a latent vector to a feature map via a fully-connected layer, followed by a sequence of ResNet blocks with upsampling to reconstruct the image.

Parameter	PolyMNIST	BIKED
Encoder		
s_0	7	8
nf	64	32
$nf_{\max}$	1024	512
Image Size	28×28	64×64
Decoder		
s_0	7	8
nf	64	32
$nf_{\max}$	512	256
Output Size	28×28	64×64

Table 8: Architecture parameters for GMRF MCVAE. Our model uses a single latent space with an extra FC layer for off-diagonal embeddings, unlike baselines that require separate joint and modality-specific spaces.

3. **Scaling for Positive Definiteness**: To preserve diagonal dominance and ensure positive definiteness, each block row i is scaled using:

$$s_i = \sum_{\substack{j=1 \\ j \neq i}}^{M} ||\tilde{\Sigma}_{i,j}|| + \delta, \quad \alpha_i = \min\big(1, \epsilon \, \frac{||\Sigma_{i,i}^{-1}||^{-1}}{s_i}\big)$$

   Here, $\delta > 0$ is a small constant to ensure numerical stability, and $\epsilon < 1$ controls off-diagonal scaling. We use $\epsilon = 0.9$ and $\delta = 10^{-6}$ in all main experiments; Table 13 reports a small sensitivity grid. We then compute the scaled off-diagonal blocks as:

$$\Sigma_{i,j} = \tilde{\Sigma}_{i,j} \cdot \sqrt{\alpha_i \alpha_j}$$

4. **Matrix Assembly**: The final covariance matrix Σ is assembled as:

$$\Sigma = (\Sigma_{i,j})_{i,j}$$

This construction ensures that Σ remains symmetric and positive definite while effectively modeling components interactions. However, the block-wise method incurs a computational overhead of $\mathcal{O}(M^2 d^3)$. By applying the construction element-wise, the complexity is reduced to $\mathcal{O}(M^2 d^2)$. Section A.5 shows that, even with this extra $\mathcal{O}(M^2 d^2)$ term, the overall training cost remains lower than that of other baselines. Nevertheless, conditional generation using Equation 6 requires each component-specific covariance $\Sigma_{j,j}$ to be self-contained, i.e., independently derivable using only encoder j. The variance construction method disrupts this independence. To address this limitation, we adopt the simplifying assumption that $\Sigma_{j,j}$ are diagonal matrices.

A.5 Complexity comparison

Throughout this section we denote by

- M: the number of components (encoders/decoders);
- d: the dimensionality of each component-specific latent;
- C: the cost (forward *and* backward FLOPs) of a single decoder pass on one mini-batch.

Per-step complexities. Three baselines considered in our experiments have the following dominant costs:

- **MMVAE / MMVAE$^+$** : M^2 cross-reconstructions $\implies \mathcal{O}(M^2\,C)$.
- **MoPoE-VAE** : enumeration of 2^M modality subsets $\implies \mathcal{O}(2^M\,C)$.
- **GMRF MCVAE (ours)** : one $\big(Md\big)^2$ covariance build plus M self-decodes

$$\mathcal{O}\big(M^2d^2 + M\,C\big).$$

After Σ is built, all cross-component generations are obtained by closed-form Gaussian conditioning; no extra decoder passes are required during training.

When is our method competitive? Because MoPoE-VAE has an exponential cost $\mathcal{O}(2^M C)$, it is already asymptotically dominated by the polynomial budgets considered here. We equal the MMVAE/MMVAE$^+$ budget when

$$M^2d^2 + M\,C \;<\; M^2\,C \;\implies\; C \;>\; \frac{M\,d^2}{M-1}, \tag{13}$$

Dataset	M	d	Threshold on C
Copula	4	2	$C > 5.3$
PolyMNIST	5	16	$C > 320$
BIKED	5	4	$C > 20$

Table 9: Numerical thresholds obtained from (13) for the three experimental settings. For reference, two fully-connected layers of size 256×256 already cost $\sim 2 \times 10^5$ FLOPs, far above any threshold listed.

Even though GMRF MCVAE incurs an $\mathcal{O}(M^2d^2)$ covariance step, realistic decoders (convolutional or large FC) comfortably exceed the thresholds in Table 9, making our approach more efficient than the 3 baselines while still enabling closed-form partial-to-full generation.

A.6 Wall-clock time per epoch (PolyMNIST)

Table 10 reports average wall-clock time per training epoch on PolyMNIST (batch size 256, three runs, single NVIDIA RTX 3060 Laptop GPU, 6 GB). Despite the overhead of covariance assembly, GMRF MCVAE is faster per epoch than all baselines in this regime because it avoids cross-modal reconstructions: conditional samples are obtained in closed form from the joint Gaussian latent.

Model	**Time / epoch (s)**
MVAE	350.6 ± 3.2
MMVAE	306.1 ± 1.7
MoPoE-VAE	313.5 ± 2.8
MMVAE+	391.5 ± 3.2
GMRF MCVAE (ours)	128.8 ± 2.1

Table 10: PolyMNIST wall-clock seconds per epoch (mean ± std, 3 runs).

A.7 Ablations and larger (M, d) timing

Table 11 varies the fraction of masked off-diagonal covariance parameters on PolyMNIST. Table 12 compares the full GMRF covariance to a diagonal variant (no cross-component blocks), and Table 13 sweeps (ϵ, δ) in Algorithm 1. Table 14 reports mean per-epoch time when duplicating modalities to increase M and d (batch size reduced to 32 where needed for memory).

Model (% mask)	FID (Uncond.)	Coh. (Uncond.)	FID (Cond.)	Coh. (Cond.)
GMRF (0%)	116.43 ± 1.20	0.290 ± 0.019	177.71 ± 0.03	0.853 ± 0.096
GMRF (50%)	118.25 ± 2.40	0.310 ± 0.020	180.23 ± 1.81	0.861 ± 0.060
GMRF (75%, default)	118.21 ± 1.71	0.321 ± 0.014	180.76 ± 3.11	0.869 ± 0.016

Table 11: PolyMNIST masking ablation (mean ± std).

Variant	FID (Uncond.)	Coh. (Uncond.)
Diagonal $\mathbf{\Sigma}$ (no cross blocks)	141.00 ± 0.59	0.12 ± 0.07

Table 12: Diagonal-covariance ablation on PolyMNIST (conditional Gaussian conditioning is not applicable in the same form).

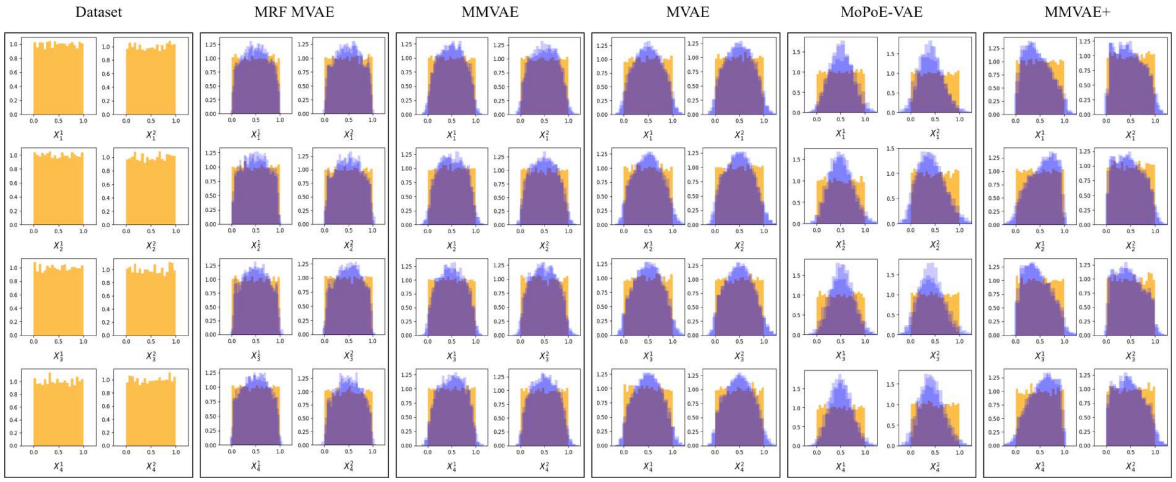

Figure 5: Qualitative analysis of unconditional generations using the Copula dataset. Each subplot displays the marginal distributions for each coordinate: $(\mathbf{x}_i^1)$ on the left and $(\mathbf{x}_i^2)$ on the right, across four two-dimensional components $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$. True distributions are depicted in orange and generated distributions in blue.

A.8 BIKED block correlations in latent space

For each test sample and training run we convert the posterior covariance into a correlation matrix, average over samples and runs, then average entries between latent dimensions belonging to each component pair. We observe consistent relative structure (which pairs are stronger than others) compatible with mechanically coupled parts, while absolute magnitudes stay moderate due to diagonal-dominance regularization ($\epsilon = 0.9$ in all main experiments).

A.9 Additional Results from the Copula Experiment

Marginal Distributions As shown in Figure 5, the marginal generations from the various baseline models and the GMRF MCVAE, generally conform to the expected range. Notably, the GMRF MCVAE closely matches the empirical marginal distributions of the dataset, consistently producing outputs within the defined range of [0,1].

Unconditional MVAE Generations Figure 6 displays the MVAE results after three independent training iterations, revealing inconsistent alignment with the actual joint distributions between components. The MVAE tends to focus selectively on certain components, often overlooking others. This behavior reflects the "veto" effect described in Shi et al. (2019), where overconfident experts disproportionately influence the model's output. Such biases negatively impact the global coherence, compromising the accurate representation of inter-component relationships.

$\delta \setminus \epsilon$	1.0	0.9	0.85
10^{-7}	118.92 ± 1.72 / 0.320 ± 0.020	118.54 ± 0.96 / 0.320 ± 0.015	119.05 ± 1.83 / 0.315 ± 0.047
10^{-6}	118.81 ± 1.70 / 0.319 ± 0.089	118.21 ± 1.71 / 0.321 ± 0.014	119.16 ± 1.11 / 0.312 ± 0.016
10^{-5}	118.97 ± 2.18 / 0.313 ± 0.048	118.63 ± 1.35 / 0.320 ± 0.013	120.11 ± 2.02 / 0.307 ± 0.058

Table 13: Sensitivity to (ϵ, δ) in Algorithm 1 (PolyMNIST: unconditional FID / coherence). Default is $\epsilon = 0.9$, $\delta = 10^{-6}$.

M	d	**Avg. epoch duration (s)**
10	32	294.0 ± 18.4
10	64	408.5 ± 25.6
10	128	1899.9 ± 418.9
15	32	466.4 ± 29.2
15	64	1017.8 ± 263.7
15	128	20789.1 ± 1301.5

Table 14: Synthetic scaling on duplicated PolyMNIST modalities (mean ± std per epoch; batch 32).

A.10 Exploring a Generalized Variant of GMRF MCVAE Models

In this section, we discuss a more comprehensive configuration in which both the prior $p(\mathbf{z})$ and posterior $p(\mathbf{z}|\mathbf{X})$ are characterized by general Markov Random Fields. This approach opens up possibilities for robustly modeling complex inter-component relationships. However, this configuration also presents significant challenges due to the intractability of the partition functions $\mathcal{Z}_p$ and $\mathcal{Z}_q$, which are critical to the prior and posterior distributions. The ELBO for this model configuration is as follows:

$$\begin{aligned} \text{ELBO} = & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}[p(X|z)] - \log\left(\frac{\mathcal{Z}_p}{\mathcal{Z}_q}\right) \\ & - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}\left[\sum_{i<j}\left(\psi^p_{i,j}(z_i, z_j) - \psi^q_{i,j}(z_i, z_j)\right)\right] \\ & - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}\left[\sum_{i}\left(\psi^p_{i}(z_i) - \psi^q_{i}(z_i)\right)\right] \end{aligned} \tag{14}$$

While direct computation of $\mathcal{Z}_p$ and $\mathcal{Z}_q$ remains elusive, we can effectively estimate the gradient of the log partition function with respect to the model parameters (θ) through sampling. This estimation can be expressed as follows (Khoshaman and Amin, 2018):

$$\nabla_\theta \ln Z_\theta = \nabla_\theta \ln \sum_z \exp\left(-E_\theta(\mathbf{z})\right) = -\mathbb{E}_{p_\theta(\mathbf{z})}\left[\nabla_\theta E_\theta(\mathbf{z})\right] \tag{15}$$

where $E_\theta(\mathbf{z}) = \sum_{i<j} \psi_{i,j}(z_i, z_j) + \sum_i \psi_i(z_i)$ represents the energy of configuration z under the model parameters θ. This approach enables us to navigate the partition function's intractability, facilitating the model's training through gradient-based optimization techniques.

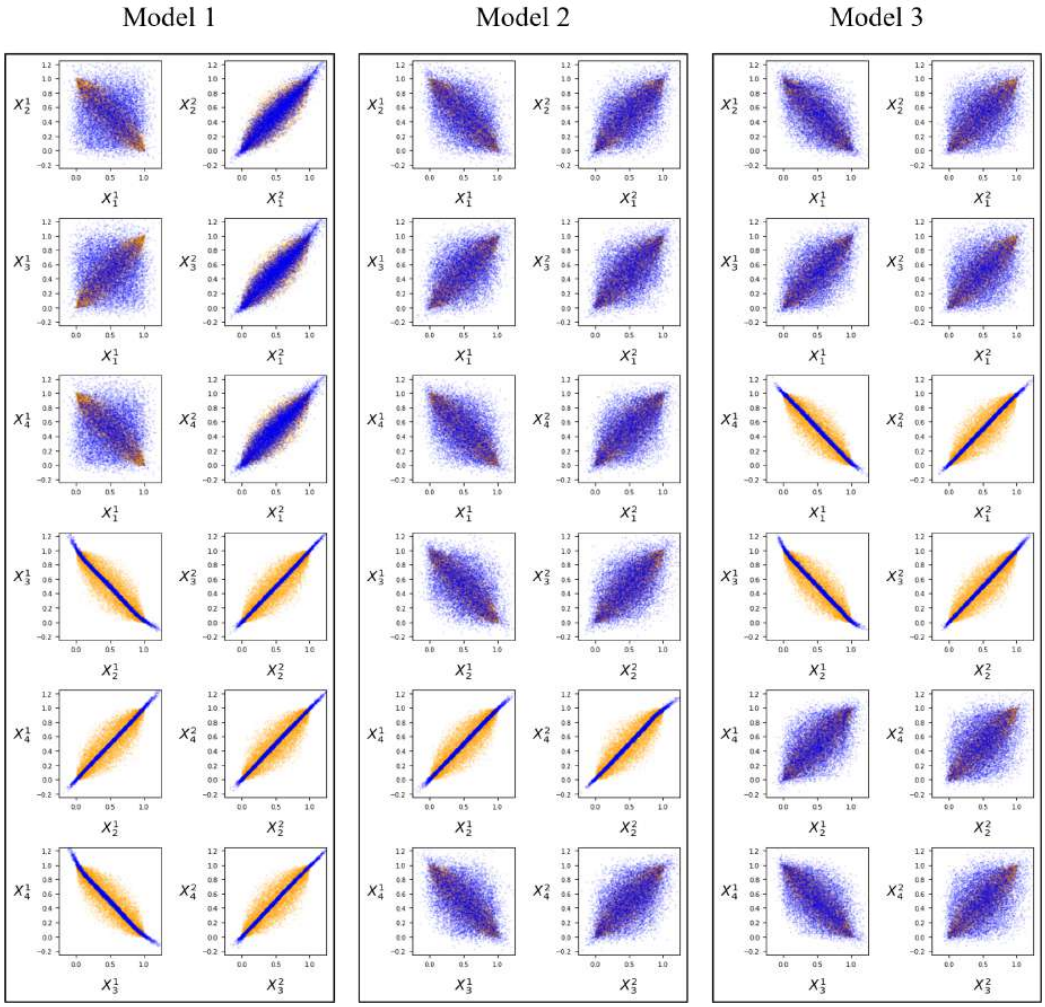

Figure 6: Qualitative results of unconditional generations from the Copula dataset across three training iterations of the MVAE. Each subplot shows joint distributions for pairs of coordinates $(\mathbf{x}_i^1, \mathbf{x}_j^1)$ and $(\mathbf{x}_i^2, \mathbf{x}_j^2)$ across the four two-dimensional components $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$. The true distributions are shown in orange, and the MVAE-generated distributions are in blue.