ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Multi Modal Generative Learning utilizing Normalizing Flows

Master Thesis

Hendrik J. Klug

Advisors: Prof. Dr. Julia Vogt, MSc. Thomas M. Sutter
Department of Computer Science, ETH Zürich

## Abstract

Multi modal, generative models are able to learn underlying generative factors of multiple data types without the need for supervision. Existing methods use a fixed, pre-selected aggregation function to merge the learned representation of each modality into a joint posterior distribution. Here, we generalise previous work by implementing the aggregation over modalities using a trainable generalized $f$-means. We show that this more flexible way to fuse the information between modalities improves the ability of the model to learn a meaningful joint posterior approximation and to generate coherent samples across data types.

# Contents

Chapter 1

# Introduction

Similar to how humans learn and extract information from their surroundings using an aggregation of their senses, a machine learning model can learn from multiple data types. Multimodal data naturally grants self-supervision in the form of shared information connecting the different data types. It also serves as an inherent regularization which forces the model to learn more robust features from the data, since these features need to be connected between modalities (Baltrušaitis, Ahuja and Morency, 2019). This may lead to more interpretable features for humans since they also infer from multiple modalities. A model that can generate any of the learned modalities, given any subset of modalities can be used for translation between modalities for example, such as image captioning. It can also find applications in the medical domain, where the model could generate, conditioned on images and medical data of a patient, a text describing the medical condition of a patient.

However, the understanding of different modalities and the interplay between data types are non-trivial research questions and longstanding goals in machine learning research (Ngiam et al., 2011). While fully supervised approaches have been applied successfully (Karpathy and Fei-Fei, 2015; Tsai et al., 2018a), the labeling of multiple data types remains time-consuming and expensive. Models that efficiently learn from multiple data types in a self-supervised fashion are much more widely applicable for real world problems. In the medical domain, for example, self-supervised training paradigms are especially useful since there labeled data is expensive to acquire and thus very scarce. Generative models represent a natural way to learn underlying generative factors of the data, in a self-supervised fashion.

Self-supervised, multi modal generative models have been applied to toy datasets (Wu and Goodman, 2018; Shi et al., 2019a; Sutter, Daunhawer and Vogt, 2020b) and real world data (Klug, Sutter and Vogt, 2021), however results have shown that current methods are not able to aggregate well enough over the modalities to generate coherent samples. For the model to generate coherent samples, it needs to extract and fuse information from the multiple data types. An image captioning model for example, needs to extract information from the image and generate text from it when generating the caption for an image of a green apple. Captions such as ”A red apple.” or ”A yellow truck.” would not be coherent with the image of a green apple.

In previous work, the aggregation over modalities is done with multiple, fixed, pre-selected methods, each coming with advantages and disadvantages. Here, we generalise previous work by implementing the aggregation over modalities using a generic function with trainable parameters. We show that this more flexible way to fuse the information between modalities improves the ability of the model to generate coherent samples across data types.

Chapter 2

# Related Work

## 2.1 Generative Modeling

Generative adversarial networks (Goodfellow et al., 2014, GANs) and variational autoencoders (VAEs, section 3.1) are the two most popular methods for generative modeling. Both attempt to model the distribution over the data, however while for the VAEs, the resulting posterior approximation is defined explicitly, the learned posterior of GANs can not be evaluated directly. GANs are made of two models that are trained simultaneously, a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The joint optimization of both models D and G can be tricky in practice and GANs are known to suffer from mode collapse since the objective does not require the learned representation to contain all modes of the data. For images of animals for example, the generator G could learn to generate only images of brown, short haired dogs, so well that the discriminator D will not be able to distinguish them from the true data. Mode collapse does not happen in VAEs since their objective explicitly requires their learned representation to contain all modes of the data. Also, since the learned posterior distribution of VAEs can be evaluated explicitly, additional constraints can be added to the objective to push the posterior distribution to have specific characteristics and it can be used for downstream tasks like clustering or classification. In this work, we focus on VAEs and give a more in depth introduction in section section 3.1.

## 2.2 Multi Modal Generative Modeling

There have been a wide range of approaches for multi-modal generative modeling, however most fall short of expressing the complete range of behaviour that we expect in this setting.

**Modality Translation** Most prior approaches to generative modelling with multi modal data have targeted modality translation, where the model learns to generate one modality conditioned on another one. In this case input an output modalities of the model are not interchangeable. Modality translation has been proposed both as VAE based (Pu et al., 2016; Pandey and Dukkipati, 2017), as well as GAN based, for domain translation of images (Ledig et al., 2017; Liu et al., 2019). However, we expect our method to be able to generate any modality given any subset of modalities which extends translation between modalities. It would be possible to train $2^M - 1$ modality translation network pairs for $M$ modalities, but this is intractable in practice.

**Joint approximation** Other prior work has targeted to directly model the joint distribution over the data. The joint multi modal VAE (JMVAE) from (Suzuki, Nakayama and Matsuo, 2016) learns a joint posterior distribution using a joint inference network. To handle missing data at test time, inference networks need to be trained for every subset of modalities. While feasible for two modalities, this setup quickly becomes intractable with more data types. Similarly, the multimodal factorisation model (MFM) from (Tsai et al., 2018b) explicitly defines a joint inference network on top of uni modal encoders, however additional decoder networks are needed to generate missing modalities.

These approaches typically do not scale well with the number of modalities since they require additional modelling components for each combination of modalities. The MVAE from (Wu and Goodman, 2018) marked an improvement over previous methods in this regard, proposing to model the joint posterior as a product of experts (POE) over the marginal posteriors, enabling cross-modal generation at test-time without requiring additional inference networks and multi-stage training regimes. Since then, other methods have emerged, each proposing another aggregation function over the marginal posteriors. We refer to section 3.2 for a more in depth introduction to the MVAE and other methods that build on it.

Next to the aggregation function with which the uni modal posteriors are merged, other methods have been proposed to improve multi modal VAEs (mmVAEs). In (Daunhawer et al., 2021), the authors propose to split the latent space into modality specific and shared information in order to disentangle (Burgess et al., 2018) them in a purely self-supervised manner. The aggregation of modalities should only happen over the shared information and thus it makes sense to separate it from the modality specific information in order to simplify the aggregation. For this, the authors add a new term to the mmVAE objective, which disentangles the shared representations with the modality specific representations and encourages mutual information between representations that contain shared information. This has been shown to improve the conditional generation of missing modalities, however the results from (Sutter, Daunhawer and Vogt, 2020a) point out that independent of that separation, the generation coherence differs between different merging functions. The goal of this work is solely to improve the merging function, which is why we forgo this method even though we expect the separation of shared and modality specific information to improve our results.

Chapter 3

# Background

Our methods build on concepts and previous work on variational autoencoders (VAEs), self-supervised multi modal generative learning paradigms and normalizing flows, which we introduce in this section.

## 3.1 Variational Autoencoder

The VAE, first introduced by (Kingma and Welling, 2014) and (Rezende, Mohamed and Wierstra, 2014), consists of an encoder network and a decoder network. In contrast to a typical auto encoder network, the VAE is trained such that its learned representation has the structure of a prior distribution. The most popular choice for a prior is the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, which we also use in this work. The latent representation being a distribution, the decoder part can generate unseen data by sampling from it. The model is trained such that it maximizes the log-likelihood of the data ($\log p(x)$) by maximizing the Evidence Lower BOund (ELBO):

$$\begin{aligned}
\log p(x) &= \log \int p(x,z) dz \\
&= \log \int p(x,z) \frac{q(z|x)}{q(z|x)} dz \\
&\geq \mathbb{E}_{q(z|x)}[\log \frac{p(x|z)p(z)}{q(z|x)}] \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathbb{E}_{q(z|x)}[\log \frac{q(z|x)}{p(z)}] \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}\left(q(z|x) \;||\; p(z)\right) \\
&= ELBO
\end{aligned} \tag{3.1}$$

The ELBO consists of two parts: the reconstruction loss which pushes the generated samples to resemble the real data and a regularization term which forces the latent representation to be structured like the prior. In (Higgins et al., 2016), the authors introduce the hyperparameter $\beta$, which allows to weight the regularization term in the VAE objective:

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \beta D_{KL}\left(q(z|x) \mid\mid p(z)\right) \tag{3.2}$$

A lower $\beta$ gives the model more freedom in learning the latent representation, while a higher $\beta$ forces the model to learn a latent distribution that is disentangled, like the prior. "Disentangled" here means that each dimension in the learned latent representation is independent of each other, and represents a latent factor that corresponds to a different attribute in the data. In images of animals for example, one dimension in the latent representation could represent the color of the fur, while another might correspond to the color of the eyes. Both the color of the fur and the color of the eyes are independent, and so should be the corresponding latent variables. A structured and disentangled representation leads to better interpretability and is widely believed to lead to better results in down-stream tasks, however this claim has been challenged in (Locatello et al., 2019) where the authors could not find evidence for it.

The $\beta$ allows to weight the trade-off introduced by the modified training objective that punishes reconstruction quality in order to encourage disentanglement within the latent representations (Burgess et al., 2018). Disentanglement is a popular objective in representation learning and has been addressed in recent works (Chen et al., 2019; Locatello et al., 2019). In this work, we also make use of $\beta$ as a hyperparameter that we adapt for each method.

## 3.2 Multi Modal VAEs

In order for the VAE model to learn a representation which captures the underlying factors of multiple modalities, several adaptations to the objective in eq. (3.1) need to be made. The first approach that scales with the number of modalities, allows for a coherent joint generation over all modalities and cross-generation across individual modalities, the MVAE, was introduced in (Wu and Goodman, 2018). The MVAE makes the assumption that the joint posterior of data containing M modalities $\mathbb{X} = \{\mathbb{X}_m\}_{m=1}^{M}$ is a product of uni modal posteriors, also called a Product-of-Experts (PoE) (Hinton, 2002):

$$p(z|\mathbb{X}_1, \ldots, \mathbb{X}_M) \propto \prod_{m=1}^{M} q(z|\mathbb{X}_m) \tag{3.3}$$

The PoE has the advantage of aggregating information across any subset of uni modal posteriors which allows for missing modalities. However, the product of experts does not train the individual inference networks and they don't learn to handle missing data at test time. To address this issue, the MVAE requires a sub-sampling of uni modal log-likelihoods, which no longer guarantees a valid lower bound on the joint log-likelihood (Wu and Goodman, 2019).

Another approach was proposed with the MMVAE in (Shi et al., 2019b), which models the joint posterior as a mixture of uni modal posteriors, i.e. a mixture of experts (MoE):

$$p(z|\mathbb{X}_1, \dots, \mathbb{X}_M) = \frac{1}{M} \sum_{m=1} q(z|\mathbb{X}_m) \tag{3.4}$$

The MoE has the advantage of optimizing each inference network individually, however it does not merge the information between posteriors since only uni modal posteriors are considered during training.

Both advantages of the PoE, which results in a good approximation of the joint distribution and the MoE which optimizes each uni modal posterior individually are combined in the MoPoE (Sutter, Daunhawer and Vogt, 2021). The MoPoE-VAE takes advantage of both methods by merging the uni modal posteriors into $2^M - 1$ subsets, which are then combined with a MoE (see fig. 3.1).

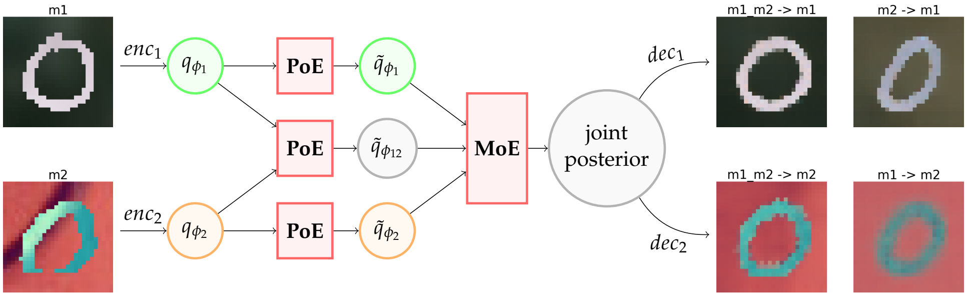


Figure 3.1: **The MoPoE makes use of the PoE to create $2^M$ subsets, which are then merged with a MoE.** Here $M = 2$, the empty subset is not shown. On the left side are the two input modalities from the polymnist dataset (see section 5.1.1), on the right side are the generated samples. In the header of each generated sample is described from which subset the decoder sampled for the generation (left side of the →) and which modality was generated (right side of the →).

Similar to the MoE, the MoPoE models the joint posterior as a mixture. However the mixture of experts consists of subsets instead of uni modal posteriors. For multi modal data $\mathbb{X} = \{\mathbb{X}_m\}_{m=1}^{M}$ with M modalities, and

$2^M - 1$ subsets of modalities $\mathbb{X}_s \in \mathbb{X}$, the objective of the MoPoE, which is an evidence lower bound (ELBO) on the joint log-likelihood $\log p_\theta(\mathbb{X})$, can be written as follows:

$$\mathcal{L}_{MoPoE}(\theta, \phi; \mathbb{X}) := \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - D_{KL}\left(\frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} \tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) \mid\mid p_\theta(\mathbf{z})\right) \tag{3.5}$$

with $q_\phi(\mathbf{z}|\mathbb{X})$ the joint posterior:

$$q_\phi(\mathbf{z}|\mathbb{X}) = \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} \tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) \tag{3.6}$$

and $\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)$ the posterior approximation of subset $\mathbb{X}_s$:

$$\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) = PoE(\{q_{\phi_m}(\mathbf{z}|\mathbb{X}_m) \forall \mathbb{X}_m \in \mathbb{X}_s\}) \propto \prod_{\mathbb{X}_m \in \mathbb{X}_s} q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)) \tag{3.7}$$

For gaussian posteriors, the PoE in eq. (3.7) can be computed in closed form. Lemma 3.1 from (Sutter, Daunhawer and Vogt, 2020a) states that the KL-divergence of the multimodal posterior distribution is a lower bound for the weighted sum of the KL-divergences of the unimodal variational approximation functions. Accordingly, Equation (3.5) can be further simplified:

$$\mathcal{L}_{MoPoE}(\theta, \phi; \mathbb{X}) \leq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} D_{KL}\left(\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) \mid\mid p_\theta(\mathbf{z})\right) \tag{3.8}$$

It has been shown in (Sutter, Daunhawer and Vogt, 2020a) that the joint generation coherence of the MoE surpasses that of the MoPoE, suggesting that a more flexible aggregation function might be needed to further improve results. In this work, the MoPoE is taken as the current state of the art for scalable, self-supervised, multi modal generative models and is used as baseline to compare our methods against. We also compare to the PoE which is seen as the gold standard for aggregating information across modalities and the MoE for learning each modality equally well and obtaining an informative joint posterior.

**Lemma 3.1 (Joint Approximation Function)** *The KL-divergence of the multimodal variational posterior approximation is a lower bound for the weighted sum of the KL-divergences of the unimodal variational approximation functions (Sutter, Daunhawer and Vogt, 2020a):*

$$D_{KL}\left(\sum_{i=1}^{M} \frac{1}{M} q_{\phi_m}(z|\mathbb{X}_m) \mid\mid p_\theta(z)\right) \leq \sum_{i=1}^{M} \frac{1}{M} D_{KL}\left(q_{\phi_m}(z|\mathbb{X}_m) \mid\mid p_\theta(z)\right) \tag{3.9}$$

## 3.3 Normalizing Flows

Normalizing flows (Papamakarios et al., 2019) represent an approach for defining invertible and differentiable transformations of probability distributions. They are widely used for generative modeling (Dinh, Sohl-Dickstein and Bengio, 2017; Kingma and Dhariwal, 2018) and variational inference (Rezende and Mohamed, 2016; Berg et al., 2019). In this work, we make use of normalizing flows both as a simple parameterizable invertible function for the $f$-mean, as well as a transformation of the joint posterior into an arbitrary complex distribution in order to improve its ability to capture the underlying factors of multiple modalities.

In practice, flow-based models are typically constructed by implementing the diffeomorphic transformation T (or $T^{-1}$) with a neural network. Because invertible and differentiable transforms are composable, complex transformations can be built by composing multiple instances of simpler ones: $T = T_F \circ \cdots \circ T_1$. The density of the transformed posterior $\tilde{q}_\phi$ can easily be obtained with the change of variable formula (Bogachev, 2007):

$$\tilde{q}_\phi = T(q_\phi) \quad \text{where} \quad q_\phi \sim p_{q_\phi}(q_\phi) = \mathcal{N}(\mu_\phi, \sigma^2_\phi) \tag{3.10}$$

$$p_{\tilde{q}_\phi}(\tilde{q}_\phi) = p_{q_\phi}(q_\phi) |\det J_T(q_\phi)|^{-1} \tag{3.11}$$

In generative modeling, normalizing flows are used to learn a diffeomorphic mapping $T$ from images to a prior, like Gaussian noise. Since $T$ is invertible, one can then transform samples from the prior into new images with $T^{-1}$.

For variational inference, normalizing flows are used to transform the posterior into a flexible, arbitrarily complex distribution by transforming it with a normalizing flow. The transformed posterior can be a much more faithfull approximation of the true underlying distribution than posterior approximations that are limited to one class, like a normal distribution with a diagonal covariance.

In this work, we make use of chained coupling blocks as normalizing flows, implemented by the Framework for Easily Invertible Architectures (FrEIA). Coupling blocks were first introduced in (Dinh, Krueger and Bengio, 2015) for their Nonlinear Independent Components Estimation (NICE) method. Chained coupling blocks are constructed such that at every block, the data is split into two halves. One half is transformed by a simple linear transformation with parameters depending on the other half. The transformed half is then concatenated with the other, unchanged half. This process is shown in fig. 3.2. This transformation has a Jacobian which determinant is easily computable since it is triangular. This gives: $\det Df(x) = \det D\hat{f}(x^B|\Theta(x^A))$. Affine coupling blocks have been shown to provide good results, especially for image data (Dinh, Sohl-Dickstein and Bengio, 2017)

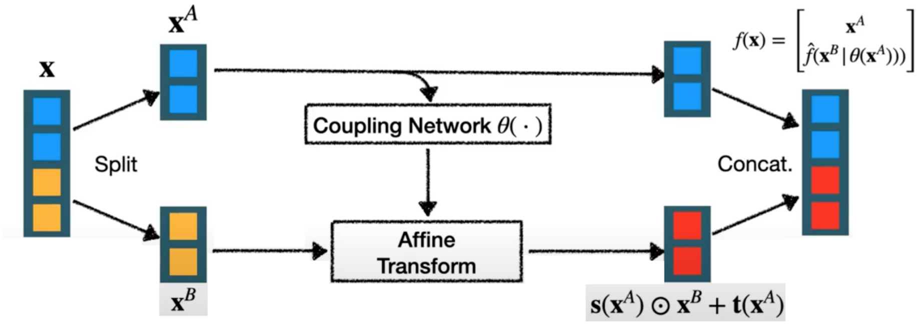

x
$x^A$
$x^B$
Split
Coupling Network θ( · )
Affine Transform
$s(x^A) \odot x^B + t(x^A)$
Concat.
$f(x) = \begin{bmatrix} x^A \\ \hat{f}(x^B \mid \theta(x^A)) \end{bmatrix}$

Figure 3.2: Flowchart depicting the workings of a coupling block, taken from the ECCV2020's Tutorial: "Introduction to Normalizing Flows" (Brubaker, 2020)

## 3.4 Importance Weighted Autoencoder

It has been shown that the tightness of the ELBO in eq. (3.2) can be improved by sampling multiple samples from the posterior at each step (Burda, Grosse and Salakhutdinov, 2016), which results in the following lower bound:

$$\log p(x) \geq \mathbb{E}_{z_1,\ldots,z_K \sim q_\phi(z|x)} \left[ \log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(x|z_k) p_\theta(z)}{q_\phi(z_k|x)} \right] := \mathcal{L}_K \tag{3.12}$$

Equation (3.12) yields useful properties summarized in (Nowozin, 2018), namely that one recovers the ELBO for $K = 1$, $\mathcal{L}_K$ approaches the true $\log p(x)$ for $K >> 1$ ($\lim_{K \to \inf} \mathcal{L}_K = \log p(x)$) and $\mathcal{L}_1, \ldots, \mathcal{L}_K$ provide stochastic monotonicity ($\mathcal{L}_E = \mathcal{L}_1 \leq \mathcal{L}_2 \leq \ldots \leq \log p(x)$). The MMVAE from (Shi et al., 2019b) adapts this for multi modal data:

$$\mathcal{L}_K^{MoE}(x_{1:M}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{z_m^{1:K} \sim q_{\phi_m}(z|x_m)} \left[ \log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(x_{1:M}|z_m^k) p_\theta(z_m^k)}{q_\phi(z_m^k|x_{1:M})} \right] \tag{3.13}$$

which is a valid lower bound of the multi modal log likelihood $\log p(\mathbb{X})$.

In our work, we make use of this importance sampling training paradigm to improve the tightness of the ELBO and to approximate the KL-divergence between the posterior and the prior (see section 4.3.4).

Chapter 4

# Methods

## 4.1 Learning a flexible joint posterior with a generalized $f$-Mean

As introduced herein, we are working with a multi modal VAE (mmVAE), which learns a joint distribution that contains the combined information of each learned uni modal latent distribution. For $M$ modalities, $M$ different encoder and decoder pairs are needed, each encoder learning a unimodal latent distribution $q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)$. To learn a joint distribution of multiple data modalities, some function $\mathcal{F}$ is needed that merges the information from all unimodal latent distributions into one joint distribution (see fig. 4.1). In previous work (Wu and Goodman, 2018; Shi et al., 2019a; Sutter, Daunhawer and Vogt, 2020b), learning a joint distribution has been done effectively by combining learned unimodal distributions with a PoE (Wu and Goodman, 2018), an MoE (Shi et al., 2019a) or both (Sutter, Daunhawer and Vogt, 2020b). While for both the MoE and PoE, reasons have been established why they are good choices for the aggregation function, both come with several shortcomings (section 3.2). A more flexible and generic function could improve the fusion of information from each modality and improve the expressiveness of the joint posterior.

To this end, we generalize previous methods and implement the fusion of the uni modal latent distributions utilizing a trainable generalized $f$-mean, with parameters $\psi$.

Since the generalized $f$-Mean is a generalisation of the arithmetic and the geometric mean, it should bring results that are at least equally good or better than previous results if the objective is right. E.g. if the geometric or the arithmetic mean were the best functions to merge the uni modal posteriors, the model could learn parameters $\psi$ such that the $f$-Mean equals an arithmetic or geometric mean.

The generalized $f$-Mean is defined as follows:

$$\mathcal{M}_f(\mathbf{p}) = f^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{p}_i))\right) \tag{4.1}$$

In eq. (4.1), $f$ can be anything as long as it is invertible and differentiable. Normalizing Flows (Papamakarios et al., 2019) present an approach to implement a sequence of invertible transformations with neural networks and provide a natural implementation for a parameterized $f_\psi$. We refer to section 3.3 for a more in-depth introduction to normalizing flows.

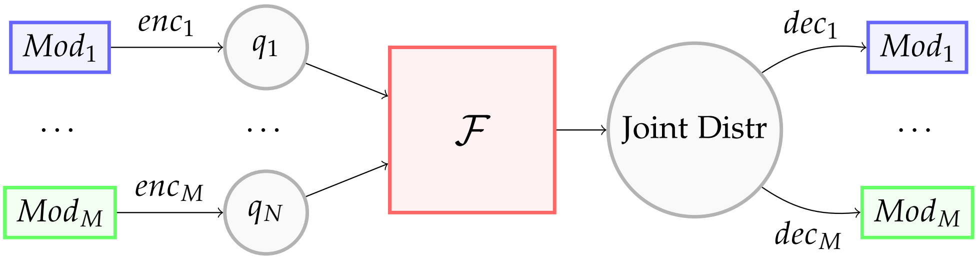


Figure 4.1: **Flowchart depicting the main elements of a multi modal VAE wit $M$ different modalities.** Each input modality $m$ gets mapped to a unimodal latent distribution $q_m$ by an encoder $enc_m$. The $M$ unimodal learned distributions then get merged by a function $\mathcal{F}$ into a joint distribution from which the decoders can sample in order to reconstruct each of the $M$ modalities.

## 4.2 Evaluating the joint posterior distribution

The main difficulty in our approach comes from the fact that the $f$-mean of the uni modal approximations follows an unknown distribution. While this makes the joint distribution more flexible, this also makes the computation of the regularization term in the ELBO, the KL-divergence, more difficult to compute. In fact, if the density of the joint distribution is unknown, it is impossible to compute the KL-divergence in closed form.

An intuitive alternative would be to find an upper bound of the KL-divergence which can be computed in closed from, such that it can be minimized in order to minimize the true divergence:

$$\begin{aligned} D'_{KL} &\geq D_{KL}(\mathcal{M}_f(\{q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)\ \forall\ \mathbb{X}_m \in \mathbb{X}_s\})) \\ &= D_{KL}\left(f^{-1}\left(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))}{|\mathbb{X}_s|}\right) \mid\mid p_\theta(\mathbf{z})\right) \end{aligned} \tag{4.2}$$

Using the change of variable formula (eq. (3.11)), the $f$-mean in eq. (4.2) can be rewritten as follows:

$$\mathcal{M}_f = f^{-1}(Q)|J_{f^{-1}}(Q)| \tag{4.3}$$

with

$$Q = \sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)|J_f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))|}{|\mathbb{X}_s|} \tag{4.4}$$

Here Q is a sum of random variables, which can be rewritten as chained convolutions (Wikipedia, 2021) and is hard to evaluate.

Instead, we propose four workarounds to the computation of the KL-divergence in eq. (4.2).

1. For one, eq. (4.2) can be simplified by skipping the backwards transformation $f^{-1}$. This leads to a mixture of transformed posteriors, which divergence can be bounded using lemma 3.1 from (Sutter, Daunhawer and Vogt, 2020a).

   We then get an upper bound that can be minimized:

$$\begin{aligned} &D_{KL}\left(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f_\psi(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))}{|\mathbb{X}_s|} \;||\; p_\theta(\mathbf{z})\right) \\ &\leq \frac{1}{|\mathbb{X}_s|} \sum_{\mathbb{X}_m \in \mathbb{X}_s} D_{KL}\left(f_\psi(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))||\; p_\theta(\mathbf{z})\right) \end{aligned} \tag{4.5}$$

   We implement this in the Mixture of flow of product of experts (MofoPoE) model, which is described in section 4.3.1.

2. Another way to simplify the KL-divergence in eq. (4.2) is to force the output of the $f$-mean to be a Gaussian distribution. This can be done by, instead of mixing the posteriors which follow a normal distribution, mixing their parameters $\mu_s$ and $\sigma_s$. The joint posterior is then described as follows:

$$q_{\phi,joint} \sim \mathcal{N}\left(f_\mu^{-1}(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f_\mu(\mu_s)}{|\mathbb{X}_s|}),\; f_\sigma^{-1}(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f_\sigma(\sigma_s^2)}{|\mathbb{X}_s|})\right) \tag{4.6}$$

   This is implemented as the mixture of parameter generalized $f$-mean (MopgfM) and described in section 4.3.2.

3. The sum of random variables in the $f$-mean (eq. (4.2)) is hard to evaluate since the transformed uni modal posteriors ($q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)$) follow an unknown distribution. It is however possible to steer the normalizing flow $f_\psi$ to map $q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)$ towards a normal distribution, such that the

sum of random variables can be evaluated. This normal distribution can be amortized by making it dependent on the input. We implement this as the MogfM_amortized method, described in section 4.3.3.

4. Instead of evaluating the density of the sum of random variables inside the $f$-mean, we investigate if it is possible to approximate it with a normal distribution. The mean and variance can be inferred using importance samples from the sum of random variables. We implement this as the importance weighted mixture of generalized $f$-mean (iwMogfM), described in section 4.3.4.

## 4.3 Models

In this section, we describe the models that implement the four methods introduced above and enumerate their advantages and disadvantages.

### 4.3.1 Mixture of flow of Products of Experts

The Mixture of flow of Products of Experts (MofoP) builds on the MoPoE by transforming the subset posterior approximations $\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)$ with a series of F invertible transformations with trainable parameters $\psi$:

$$z_{F,S} = f_\psi(z_{0,S} \sim \tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)) = f_F \circ \ldots \circ f_2 \circ f_1(z_{0,S} \sim \tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)) \tag{4.7}$$

The density of the resulting transformed subset posterior distribution can be evaluated with the change of variables formula (eq. (3.11)):

$$\ln f(\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)) = \ln q_\phi(z_0|\mathbb{X}_s) - \sum_{i=1}^{F} \ln \left| \det \frac{df_i}{dz_{i-1}} \right| \tag{4.8}$$

Here $f(\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s))$ can follow any distribution and is thus more flexible than the gaussian subset posterior approximation in the MoPoE model. A flow chart depiction of the MofoP is shown in fig. 4.2. During a forward pass, a sample is taken from each subset posterior distribution, transformed with a normalizing flow $f$ and then mixed with a MoE.

The resulting objective can be written as follows, by slightly modifying the MoPoE objective from eq. (3.8):

$$
\begin{aligned}
&\mathcal{L}_{MofoP}(\theta, \phi, \psi; \mathbb{X}) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} D_{KL}\Big(\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) \;||\; p_\theta(\mathbf{z})\Big) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} D_{KL}\Big(f_\psi(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))) \;||\; p_\theta(\mathbf{z})\Big)
\end{aligned}
\tag{4.9}
$$

The KL-divergence between the transformed subset posteriors and the prior can be evaluated as follows using eq. (4.8):

$$
\begin{aligned}
&D_{KL}\Big(f_\psi(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))) \;||\; p_\theta(\mathbf{z})\Big) \\
&= \mathbb{E}_{f_\psi(q_{\phi_m})}\left[\log f_\psi(q_{\phi_m}(f_\psi(z)|x_m)) - \log p_\theta(f_\psi(z))\right] \\
&= \mathbb{E}_{q_{\phi_m}}\Big[\log q_{\phi_m}(z|x_m) - \log\det J_{f_\psi} - \log p_\theta(f_\psi(z))\Big]
\end{aligned}
\tag{4.10}
$$

We use the MofoP method in comparison to the other methods that make use of the inverse transform $f^{-1}$ to evaluate if the merging of information between unimodal posteriors can be improved by simply making the subsets more flexible.

One advantage of the MofoP method is that since the inverse of the flow transformation is not needed, implementations of normalizing flows can be used where the evaluation of the inverse flow does not need to be tractable. This gives more flexibility in the choice of the flow implementation.

### 4.3.2 Mixture of parameter generalized $f$-means

The Mixture of parameter generalized $f$-mean (Mopgfm) mixes the means and the standard deviations of the unimodal posteriors, in order to obtain a normal distribution that depends on each of the uni modal posteriors (see eq. (4.6)). The aggregation over the means and the standard deviations is done with a parameterized $f$-mean.

This is a generalisation of the PoE method since a product of gaussian experts is itself Gaussian with mean $\mu_{PoE} = (\sum_i \mu_i V_i)(\sum_i V_i)^{-1}$ and covariance $V_{PoE} = (\sum_i V_i)^{-1}$ where $\mu_i, V_i$ are the parameters of the $i$-th Gaussian.
Without loss of generality, it can be assumed that:
$f_\mu^{-1}(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f_\mu(\mu_s)}{|\mathbb{X}_s|}) = \mu_{PoE}$ and $f_\sigma^{-1}(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f_\sigma(\sigma_s^2)}{|\mathbb{X}_s|}) = V_{PoE}$.

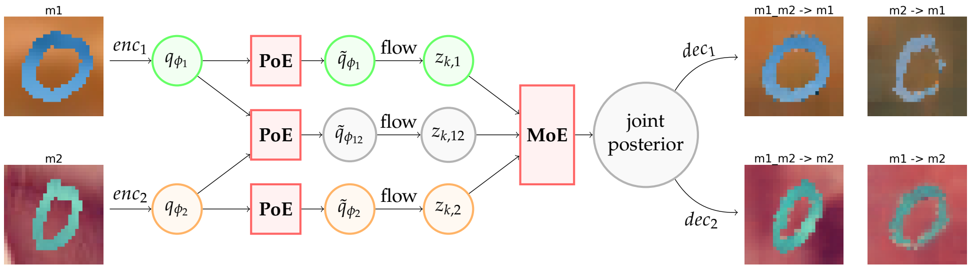


Figure 4.2: **Fowchart depicting the MofoP method.** The MofoP creates more expressive subset posteriors by transforming the PoE posteriors with a series of invertible transformations. Here, an example with 2 subsets is shown. On the left side are the two input modalities from the polymnist dataset (see section 5.1.1), on the right side are the generated samples. In the header of each generated sample is described from which subset the decoder sampled for the generation (left side of the →) and which modality was generated (right side of the →).

The main advantage of this method is that since it is a generalisation of the PoE, it gives more flexibility to the modality fusion. However, this comes at the cost that the expressiveness of the joint distribution is limited by being a Gaussian, and since the transformations are applied on the parameters of the uni modal distributions, transparency of the resulting transformation is lost. It is hard, if not impossible, to translate eq. (4.6) into the following equation:

$$q_{\phi,joint} = T(\{q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)\forall\mathbb{X}_m \in \mathbb{X}\}) \tag{4.11}$$

with T a well defined transformation. The internal workings of the MopgfM method are depicted in a simplified manner in fig. 4.3.

### 4.3.3 Amortized Mixture of generalized $f$-means

For the Amortized Mixture of generalized $f$-means (MogfM_amortized) method, we introduce a new loss $\mathcal{L}_2$ that pushes $f_\psi$ to map the uni modal posteriors to an amortized prior distribution, i.e. such that:

$$f_\psi(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)) \sim \mathcal{N}(f_\psi(\mu_m), \mathbf{I}) \tag{4.12}$$

Then, the density of the sum of random variables $\mathbf{G}_f$ can easily be evaluated with:

$$\mathbf{G}_f(\mathbf{z}|\mathbb{X}_{1:|\mathbb{X}_s|}) = \sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))}{|\mathbb{X}_s|} \sim \mathcal{N}\left(\sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f(\mu_m)}{|\mathbb{X}_s|}, \frac{1}{\sqrt{|\mathbb{X}_s|}} \cdot \mathbf{I}\right) \tag{4.13}$$

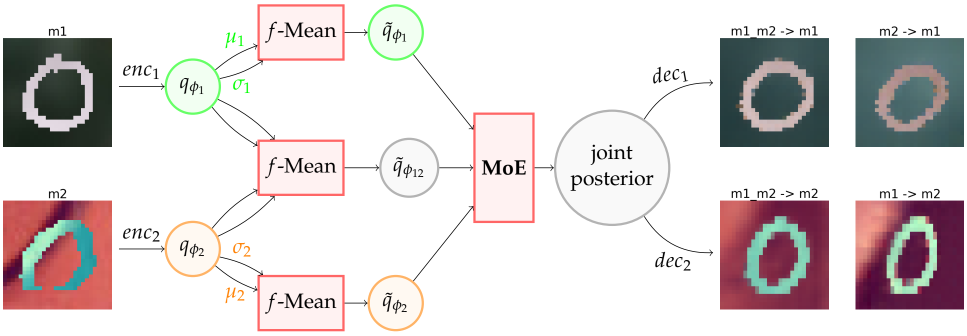


Figure 4.3: **The MopgfM makes use of the $f$-mean to aggregate the means und standard deviations of the unimodal posteriors create $2^M$ normally distributed subsets, which are then merged with a MoE.** Here, an example with 2 subsets is shown. On the left side are the two input modalities from the polymnist dataset (see section 5.1.1), on the right side are the generated samples. In the header of each generated sample is described from which subset the decoder sampled for the generation (left side of the →) and which modality was generated (right side of the →).

Equation (4.12) can be achieved by minimizing the KL-divergence between the transformed uni modal posteriors and the amortized prior:

$$\begin{aligned}
\mathcal{L}_2 &= \sum_{\mathbb{X}_m \in \mathbb{X}} D_{KL}\left(f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)) \;||\; \mathcal{N}(f(\mu_m), \mathbf{I})\right) \\
&= \sum_{\mathbb{X}_m \in \mathbb{X}} D_{KL}\left(f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)) \;||\; p_{\theta_m}(\mathbf{z})\right) \\
&= \sum_{\mathbb{X}_m \in \mathbb{X}} \mathbb{E}_{f(q_{\phi m}(\mathbf{z}|\mathbb{X}_m))}[\log f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)) - \log p_{\theta_m}(\mathbf{z})] \\
&= \sum_{\mathbb{X}_m \in \mathbb{X}} \mathbb{E}_{z_m \sim q_{\phi m}(\mathbf{z}|\mathbb{X}_m)}[\log q_{\phi_m}(z_m|\mathbb{X}_m) - \log\det J_f - \log p_{\theta_m}(f(z_m))]
\end{aligned} \tag{4.14}$$

The ELBO can then be evaluated as following:

$$\begin{aligned}
\mathcal{L}_1 &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} D_{KL}\left(\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) \;||\; p_\theta(\mathbf{z})\right) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} \mathbb{E}_{\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}[\log \tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) - \log p_\theta(\mathbf{z})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} \mathbb{E}_{\mathbf{G}_f(\mathbf{z}|\mathbf{x}_{1:|\mathbb{X}_s|})}[\log \mathbf{G}_f(\mathbf{z}|\mathbb{X}_{1:|\mathbb{X}_s|}) \\
&\quad + \log\det J_{f^{-1}} - \log p_\theta(\mathbf{z})]
\end{aligned} \tag{4.15}$$

The total loss is then:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_1 + \mathcal{L}_2 \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}[\log(p_\theta(\mathbb{X}|\mathbf{z}))] - \frac{1}{2^M} \sum_{\mathbb{X}_s \in \mathcal{P}(\mathbb{X})} \mathbb{E}_{\mathbf{G}_f(\mathbf{z}|\mathbf{x}_{1:|\mathbb{X}_s|})}[\log \mathbf{G}_f(\mathbf{z}|\mathbb{X}_{1:|\mathbb{X}_s|}) \\ &\quad + \log \det J_{f^{-1}} - \log p_\theta(\mathbf{z})] \\ &\quad + \sum_{\mathbb{X}_m \in \mathbb{X}} \mathbb{E}_{z_m \sim q_{\phi_m}(\mathbf{z}|\mathbb{X}_m)}[\log q_{\phi_m}(z_m|\mathbb{X}_m) - \log \det J_f - \log p_{\theta_m}(f(z_m))]\end{aligned} \tag{4.16}$$

The resulting joint posterior $\mathbf{G}_f$:

$$\mathbf{G}_f(\mathbf{z}|\mathbb{X}_{1:|\mathbb{X}_s|}) = \sum_{\mathbb{X}_m \in \mathbb{X}_s} \frac{f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))}{|\mathbb{X}_s|} \tag{4.17}$$

of the MogfM_amortized method can follow any distribution and can thus be more expressive than the joint posterior in the MopgfM or MoPoE methods. The main disadvantage of this method is that the KL-divergence term in $\mathcal{L}_1$ can only be evaluated when the flow $f$ has already learned to map the uni modal posteriors towards the amortized priors.

### 4.3.4 Importance Weighted Mixture of generalized $f$-means

The central limit theorem states that a sum of independent random variables tends towards a normal distribution, even if the original variables themselves are not normally distributed. With the Importance Weighted Mixture of generalized $f$-means (iwMogfM) method, we evaluate if the sum of unimodal posteriors from eq. (4.2) can be approximated with a normal distribution. It is important to note that since the unimodal posteriors should contain shared information they are assumed to be dependent such that the independence condition for the central limit theorem is not met. We find however that the normal distribution with inferred parameters is a useful proxy which allows to evaluate the KL-divergence term in the objective from eq. (3.1). To infer the parameters of the normal distribution, we take K importance samples from the sum of unimodal posteriors and evaluate their average and variance. Importance sampling from the posterior has been done before for the iwVAE in (Burda, Grosse and Salakhutdinov, 2016) (see section 3.4).

Like the MoPoE, the iwMogfM creates the joint posterior by creating $2^M - 1$ subsets from the uni modal posteriors and then mixes them with a mixture of experts. However, instead of using a PoE to create the subsets, it uses an $f$-mean. To derive the resulting objective, we rewrite the objective from the

MoPoE for K importance samples in a first step:

$$\begin{aligned}
\mathcal{L}_1^{mopoe} &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbb{X})}\left[\log \frac{p_\theta(\mathbb{X},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbb{X})}\right] \\
&= \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log \frac{p_\theta(\mathbb{X}_s,\mathbf{z})}{\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\right] \\
&= \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{z_s\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log \frac{p_\theta(\mathbb{X}_s,\mathbf{z}_s)}{\tilde{q}_\phi(\mathbf{z}_s|\mathbb{X}_s)}\right] \\
&\leq \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{z_s^{1:K}\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(\mathbb{X}_s,\mathbf{z}_s^k)}{\tilde{q}_\phi(\mathbf{z}_s^k|\mathbb{X}_s)}\right] = \mathcal{L}_K^{mopoe}
\end{aligned} \tag{4.18}$$

Using Jensens inequality, $\mathcal{L}_K^{mopoe}$ can be rewritten as follows:

$$\begin{aligned}
&\frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{z_s^{1:K}\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(\mathbb{X}_s,\mathbf{z}_s^k)}{\tilde{q}_\phi(\mathbf{z}_s^k|\mathbb{X}_s)}\right] \\
&\geq \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{z_s^{1:K}\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\frac{1}{K}\sum_{k=1}^{K} \log \frac{p_\theta(\mathbb{X}_s,\mathbf{z}_s^k)}{\tilde{q}_\phi(\mathbf{z}_s^k|\mathbb{X}_s)}\right] \\
&= \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathbb{E}_{z_s^{1:K}\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\frac{1}{K}\sum_{k=1}^{K} \log p_\theta(\mathbb{X}_s|\mathbf{z}_s^k) - \log \frac{\tilde{q}_\phi(\mathbf{z}_s^k|\mathbb{X}_s)}{p_\theta(\mathbf{z}_s^k)}\right] \\
&= \frac{1}{|\mathcal{P}(\mathbb{X})|}\sum_{\mathbb{X}_s\in\mathcal{P}(\mathbb{X})} \mathcal{R}_s^{1:K} - D_s^{1:K}
\end{aligned} \tag{4.19}$$

where $\mathcal{R}$ is the reconstruction loss and D the KL-divergence between the subset posterior and the prior. The subset posteriors are obtained with an $f$-mean of the uni modal posteriors:

$$\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s) = f^{-1}\left(\sum_{\mathbb{X}_m\in\mathbb{X}_s} \frac{f(q_{\phi_m}(\mathbf{z}|\mathbb{X}_m))}{|\mathbb{X}_s|}\right) = f^{-1}(Q_s) \tag{4.20}$$

Since the density of $Q_S$ is hard to evaluate, $D_s^{1:K}$ is calculated with a normally distributed proxy $\tilde{Q}_s$. The mean and the variance of $\tilde{Q}_s$ are inferred from the mean and the variance of the K importance samples $\mathbf{z}_{Q_s}^{1:K} \sim Q_s$

$D^{1:K}$ is then approximated with:

$$\begin{aligned}
D_s^{1:K} &\approx \mathbb{E}_{z_s\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log \tilde{q}_\phi(\mathbf{z}_s|\mathbb{X}_s) - \log p_\theta(\mathbf{z}_s)\right] \\
&= \mathbb{E}_{z_s\sim\tilde{q}_\phi(\mathbf{z}|\mathbb{X}_s)}\left[\log Q_s + \log\det J_{f^{-1}} - \log p_\theta(\mathbf{z}_s)\right]
\end{aligned} \tag{4.21}$$

The sampling from the uni modal posteriors is depicted in fig. 4.4.

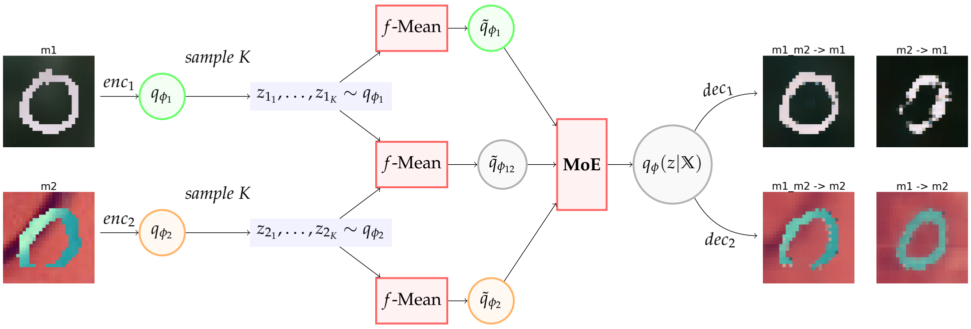


Figure 4.4: **The iwMogfM makes use of the $f$-mean to create $2^M$ subsets, which are then merged with a MoE.** Here $M = 2$, the empty subset is not shown. On the left side are the two input modalities from the polymnist dataset (see section 5.1.1), on the right side are the generated samples. In the header of each generated sample is described from which subset the decoder sampled for the generation (left side of the →) and which modality was generated (right side of the →).

Chapter 5

# Experiments

In this section we describe the experimental setup that was used in order to compare our methods to each other as well as to the MVAE, the MMVAE and the MoPoE methods.

## 5.1 Datasets

We evaluate on three datasets, each providing different difficulties in order to filter out advantages and disadvantages of our methods.

### 5.1.1 PolyMNIST

The PolyMNIST dataset, first introduced in (Sutter, Daunhawer and Vogt, 2020b), consists of MNIST digits overlayed over a random part of a certain background image. The modality specific information of each sample in this dataset is defined by the background image and the shared information by the digit. In this case the modality specific information is harder to learn than the shared information (for the modality specific information the model has to have learned the set of possible backgrounds and styles of handwriting while the shared information is simply the set of digits). Examples from the PolyMNIST dataset are shown in fig. 5.1. In total there are 60,000 tuples of training examples and 10,000 tuples of test examples. The PolyMNIST dataset is useful to study how the number of modalities impacts the performance of multi modal methods, since an abritrary amount of modalities can easily be generated.

### 5.1.2 MIMIC-CXR Database

The MIMIC-CXR Database (Johnson et al., 2019) is a large publicly available dataset of chest radiographs with free-text radiology reports containing 377,110 images corresponding to 227,835 radiographic studies performed at

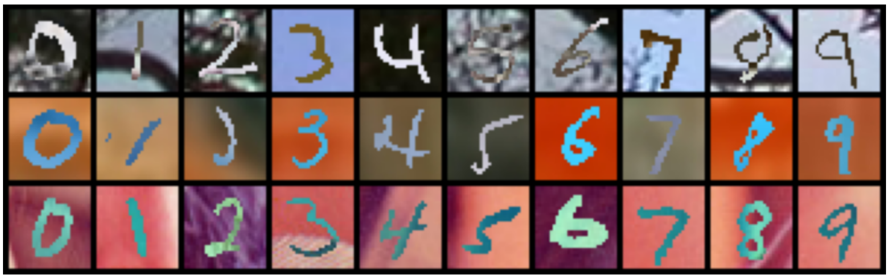
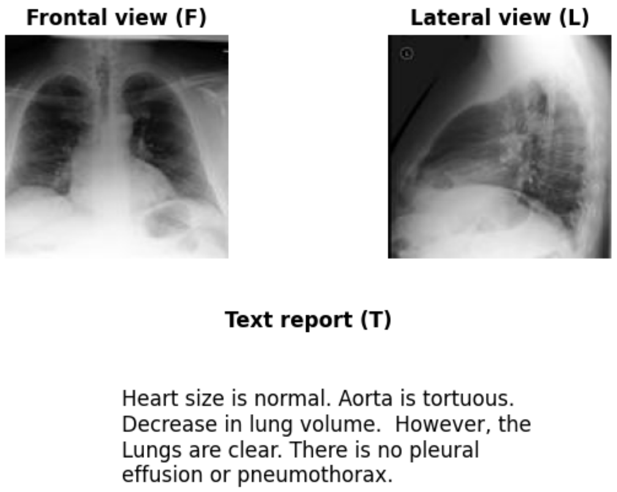


Figure 5.1: The PolyMNIST dataset consists of sets of MNIST digits where each set consists of M images with the same digit label but different backgrounds and different styles of hand writing for M different modalities.

the Beth Israel Deaconess Medical Center in Boston, MA. In this work, three modalities were extracted from the database: frontal and lateral chest radiographs together with their corresponding text reports (fig. 5.2). Only datapoints where all three modalities are present were selected. Every sample is labeled with one or more of the following categories: 'Atelectasis', 'Cardiomegaly', 'Consolidation', 'Edema', 'Enlarged Cardiomediastinum', 'Fracture', 'Lung Lesion', 'Lung Opacity', 'Pleural Effusion', 'Pleural Other', 'Pneumonia', 'Pneumothorax', 'Support Devices'. For our purposes, all images were resized to (128, 128).

**Text preprocessing** Every word that occurs at least 3 times in all the text reports is mapped to an index. Using this mapping each sentence is encoded into a sequence of indices. All sentences with a word count above 128 are truncated and all sentences consisting of less words are padded with a padding token "$< pad >$" such that all text samples are of equal length (128 words).

The MIMIC-CXR dataset is extremely challenging since both the modality specific and shared information present small details that are hard to learn. In particular, the pathologies represent only a small fraction of the images such that they are hard to distinguish, even for human experts. Also, the shared information between modalities is different between the image modalities and the image and text modalities together. The shared information between images contains information about the patient such as the posture and size, that is not contained in the text modality. The MIMIC-CXR dataset provides a good representation of real world data with all the challenges that come with it, such as unevenly represented classes and different shared information between modalities.



Heart size is normal. Aorta is tortuous. Decrease in lung volume. However, the Lungs are clear. There is no pleural effusion or pneumothorax.

Figure 5.2: An example from the MIMIC-CXR dataset is shown: the frontal view image together with the corresponding lateral view image and the text report.

## 5.2 Metrics

In order to compare the proposed methods in a meaningful manner, we make use of three metrics that each quantifies the performance of a different aspect of mmVAEs. Namely, we compare the quality of the learned latent representation, the coherence of the generated samples and the quality of the generated samples, as described in the follwing sections.

### 5.2.1 Evaluation of the Latent Representation

To evaluate if the different mmVAEs are able to extract characteristic information and compress it in the latent representation in a meaningful manner, we evaluate the separability of the latent space via linear classifiers. If the classifier can separate the latent space into the corresponding classes, we conclude that the posterior approximations are meaningful. One classifier for each class and for each latent space is trained on 1000 encoded samples from the training set and tested on the test set. Note that this can be seen as a variant of the disentanglement metric from (Higgins et al., 2016) where each class is a different generative factor. If the dimensions of latent representation are independent and interpretable, there will be less variance in the samples belonging to the same class and thus make them separable from the rest with low capacity classifiers. It has been shown in (Locatello et al., 2019) that this disentanglement metric correlates with other disentanglement metrics.

### 5.2.2 Evaluation of the Generation Coherence

To evaluate if the method is able to separate the shared information from the modality specific information, we verify that all generated tuples belong to the same class using pretrained classifiers. For conditional generation, the conditionally generated samples have to be coherent to the input samples. The coherence accuracy is the ratio of coherent samples divided by the number of generated samples. For every data type, we train a neural network classifier in a supervised way and the architecture is identical to the encoder except from the last layer.

When comparing the coherence accuracy for methods trained on only one modality, the coherence is evaluated by verifying if the generated sample belongs to the same class than the input sample. We compare the coherence accuracy for the generation of missing modalities, reconstruction of modalities and randomly generated samples.

### 5.2.3 Evaluation of the Generation Quality

To evaluate the quality of the generated samples, we make use of the precision-recall score from (Sajjadi et al., 2018). The Precision and Recall for Disitributions (prd) metric is similar to the Fréchet Inception Distance (FID) (Heusel et al., 2017), but disentangles the quality of generated samples (precision) from the coverage of the target distribution (recall). The prd metric reduces the problem of comparing a distribution Q (the distribution of generated samples) to a reference distribution P (the distribution of true images) into a one dimensional problem by applying a pre-trained classifier trained on natural images and to compare $\hat{\mathrm{P}}$ and $\hat{\mathrm{Q}}$ at a feature level. The embeddings are then clustered such that the histogram over the cluster assignments can be meaningfully compared. Failing to produce samples from a cluster with many samples from the true distribution will hurt recall, and producing samples in clusters without many real samples will hurt precision (Sajjadi et al., 2018). Here we compute the prd score by taking the area under the precision-recall curve.

## 5.3 Comparison across different number of importance samples

As introduced in section 3.4, the tightness of the ELBO in eq. (3.2) can be improved by sampling multiple importance samples from the posterior at each step (Burda, Grosse and Salakhutdinov, 2016). To test if the advantage of our more flexible aggregation over modalities using the generalized $f$-mean can be overcome by taking more importance samples, we compare the mopoe and the mopgfm methods using the importance weighted training paradigm from (Burda, Grosse and Salakhutdinov, 2016), with different number of importance samples. The results are shown in section 6.2.4.

## 5.4 Hyperparameter Selection

We select three hyperparameters for the standard mmVAE models (MoPoE, MoE, PoE) that we optimize for our experiments:

- The dimension of the latent representation (the bottleneck of the VAE). A higher dimensional latent representation gives the model more freedom to separate the different classes and can contain more information in general. However, for a too large latent representation, the encoder is not constrained to extract only the most informative features of the input such that the latent representation will contain much information that is non-informative for the decoder.
- The learning rate for the stochastic optimization of the parameters, using the Adam optimizer (Kingma and Ba, 2017). For a low learning rate, the objective will take a very long time to converge and for a too high learning rate it might oscillate around a local minimum and never converge.
- The $\beta$ in the modified ELBO from eq. (3.2), described in section 3.1

Since the choice for these parameters is non trivial, we optimize them using the hyperparameter optimization framework `Optuna` (Akiba et al., 2019). As objective, we use a weighted average of the generation coherence metric (section 5.2.2) and the area Under the precision-recall curve (prd-score, section 5.2.3, where a higher weight is given to the prd-score since its values are generally lower than those of the generation coherence metric. The results for the MoPoE method can be seen in fig. 6.1.

For our methods that make use of normalizing flows, we add three additional hyperparameters:

- The number of chained transformations with which the normalizing flow is constructed (Nbr Flows).
- The number of coupling block layers per transformation (Nbr Coupling Block layers).
- The number of parameters of each coupling block layer Coupling Dim).

For the optimization of those, we fixed the dimension of the latent representation and the learning rate according to what gave the best results for the MoPoE method. Namely, a latent representation of dimension 1280 and a learning rate of $5e-4$. The results can be seen in fig. 6.2 and fig. 6.3.

| Method | Class Dim | Coupling Dim | End Epoch | Learning Rate | Nbr Coupling Block layers | Nbr Flows | beta |
|---|---|---|---|---|---|---|---|
| moe | 1280 | | 500 | 0.0005 | | | 2.0 |
| mopgfm | 1280 | 64 | 500 | 0.0005 | 8 | 1 | 2.0 |
| poe | 512 | | 500 | 0.0005 | | | 2.0 |
| mogfm_amortized | 1280 | 512 | 500 | 0.0005 | 2 | 4 | 0.0 |
| mofop | 1280 | 64 | 500 | 0.0005 | 8 | 1 | 2.0 |
| iwmogfm2 | 1280 | 512 | 500 | 0.0005 | 2 | 4 | 0.0 |
| mopoe | 1280 | | 500 | 0.0005 | | | 2.0 |

Table 5.1: Parameters used for the models evaluated on the PolyMNIST dataset.

## 5.5 General Setup

### 5.5.1 PolyMNIST

We present our results for the PolyMNIST dataset in section 6.2, for which we trained each method 2 times for 500 epochs. All results are presented as averages over the 2 runs, accompanied with the standard deviations. If the number of modalities is not explicitely specified, the model was trained with three modalities from the PolyMNIST dataset. All parameter values for the experiments on the PolyMNIST dataset studied in section 6.2 can be found in table 5.1. To reduce training time we chose to use a small number of chained transformations (Nbr Flows) for all normalizing flow methods. A lower number of flows also yielded more stable results. We adapted the parameters of the PoE and the MoE method to match those selected for the MoPoE using the hyperoptimization. Only the dimension of the latent representation (class dim) of the PoE was reduced since the performance of the PoE dropped significantly with a higher dimension. All methods are trained with one importance sample from the joint posterior if not specified otherwise, except for the iwmogfm, which is trained with two importance samples (section 4.3.4). We use the same network architecture that was used in (Sutter, Daunhawer and Vogt, 2020b), a simple 3 layer convolutional network as encoder and decoder.

All parameters for the iwmogfm and mogfm_amortized methods have been chosen without any hyperoptimization. For both methods, we found it difficult to find the optimal $\beta$, but found that both are able to learn meaningful representations and yield good generative results, without the KL-divergence as regularisation. We set $\beta$ to 0 for the mogfm_amortized and to 0.001 for iwmogfm. A very low $\beta$ yielded better results for the iwmogfm method than $\beta = 0$.

### 5.5.2 MIMIC-CXR

We present our results for the MIMIC-CXR dataset in section 6.3, for which each method was trained once for 150 epochs. All parameter for the methods evaluated on the MIMIC-CXR dataset were selected to match those used in

| | Class Dim | Coupling Dim | End Epoch | Learning Rate | Nbr Coupling Block layers | Nbr Flows | beta |
|---|---|---|---|---|---|---|---|
| Method | | | | | | | |
| mopoe | 512 | | 149 | 0.0005 | | | 1.0 |
| mofop | 512 | 64 | 149 | 0.0005 | 8 | 3 | 1.0 |
| mopgfm | 512 | 64 | 149 | 0.0005 | 8 | 3 | 1.0 |

Table 5.2: Parameters used for the models evaluated on the MIMIC-CXR dataset.

(Klug, Sutter and Vogt, 2021). We chose to weight every modality equally in the reconstruction loss. A table with all parameters for every method evaluated on the MIMIC-CXR dataset is shown in table 5.2. We use the same ResNet (He et al., 2016) type architecture for the encoder and decoder, with 5 residual layers for the image modalities and 6 residual layers for the text modality. We refer to the published codebase for more details on the implementation of the models (MMVAE_Hub, 2021).

## 5.6 Reproducibility

Advances in scientific research are contingent on reproducibility and verifiability of previous work. To this end, we make the framework used to train all models evaluated in this work available as an open source python package (MMVAE_Hub, 2021), tested with continuous integration using (Travis, 2011) and kept up to date with (Dependabot, 2020). We publish this thesis as a reproducible self publishing document (Ioanas and Rudin, 2018, RepSeP) made available on GitHub (Klug, 2021b). All data used to produce this document, including the trained models are made available on Zenodo (Klug, 2021a). Using LaTeX and PythonTeX (Poore, 2015), we make all steps described herein easily reexecutable and extendable. It is thus easy to reproduce all figures using different parameters for each method for future work.

Chapter 6

# Results

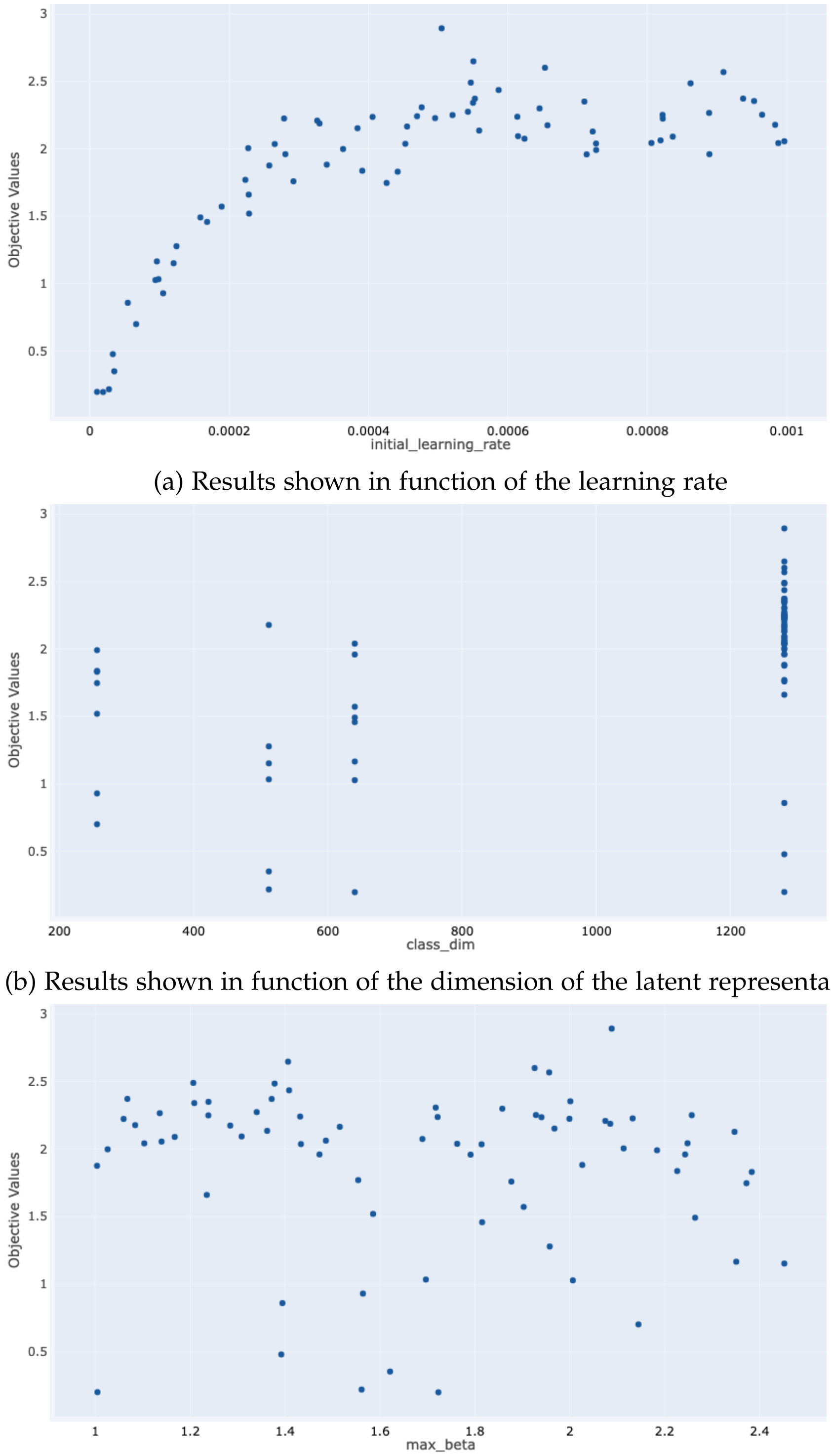
## 6.1 Hyperoptimization Results

The results for the optimization of the hyperparameters described in section 5.4 can be seen in fig. 6.1 for the MoPoE method and in fig. 6.2 and fig. 6.2 for the Mopgfm method. Note that every figure in fig. 6.1, fig. 6.2 and fig. 6.3 represents results in function of a parameter, however all other parameters are not fixed and might vary for every point.

**MoPoE Results** Descriptively, we find that the MoPoE performs best on the PolyMNIST dataset with a learning rate $\approx 5e-4$ and a latent dimension of 1280. The performance of the MoPoE seems to be robust to a change of $\beta$ in the range of 1.1 to 2.1.

**MopgfM Results** The optimal number of coupling layers appears to be 8 with the best number of dimensions being 64. Figure 6.3a shows that better scores are achieved with a higher number of chained transformations, however more flow transformations also lead to more variance in the resulting score. In practice, we have also experienced that models with a high number of normalizing flows can provide better performance but are more unstable. The Mopgfm seems to perform best with a $\beta$ between 1.5 and 2.4.

Overall the hyperoptimization results show that while the MoPoE presents results that are much more stable (from fig. 6.1a, one can infer that the only true variance in the objective value is due to a high learning rate), the highest achieved scores are lower than those achieved by the Mopgfm method.

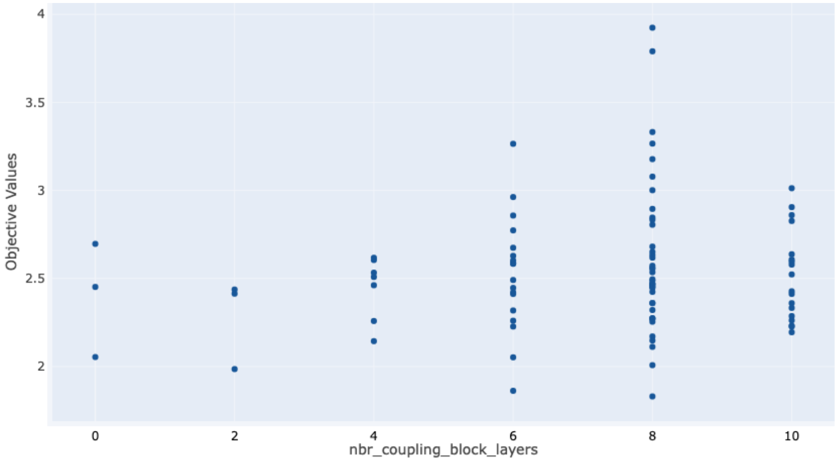


(a) Results shown in function of the learning rate

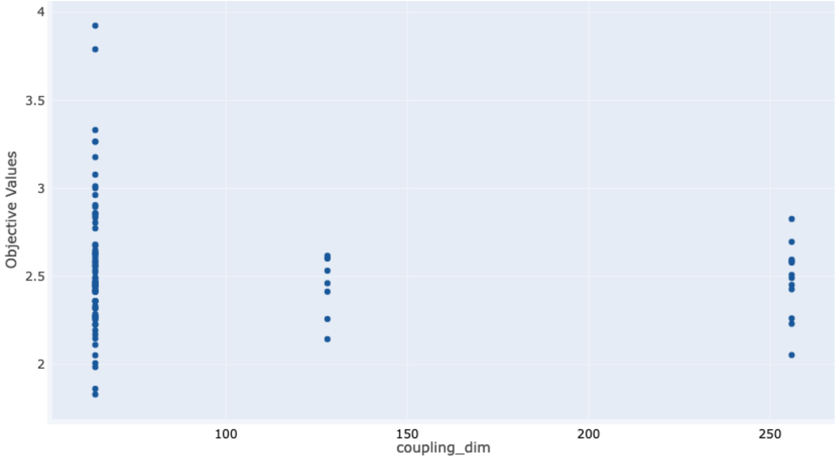
(b) Results shown in function of the dimension of the latent representation

(c) Results shown in function of $\beta$

Figure 6.1: Hyperoptimization run results for the MoPoE method. Every subfigure presents results in function of one parameter, with all other parameters varying.

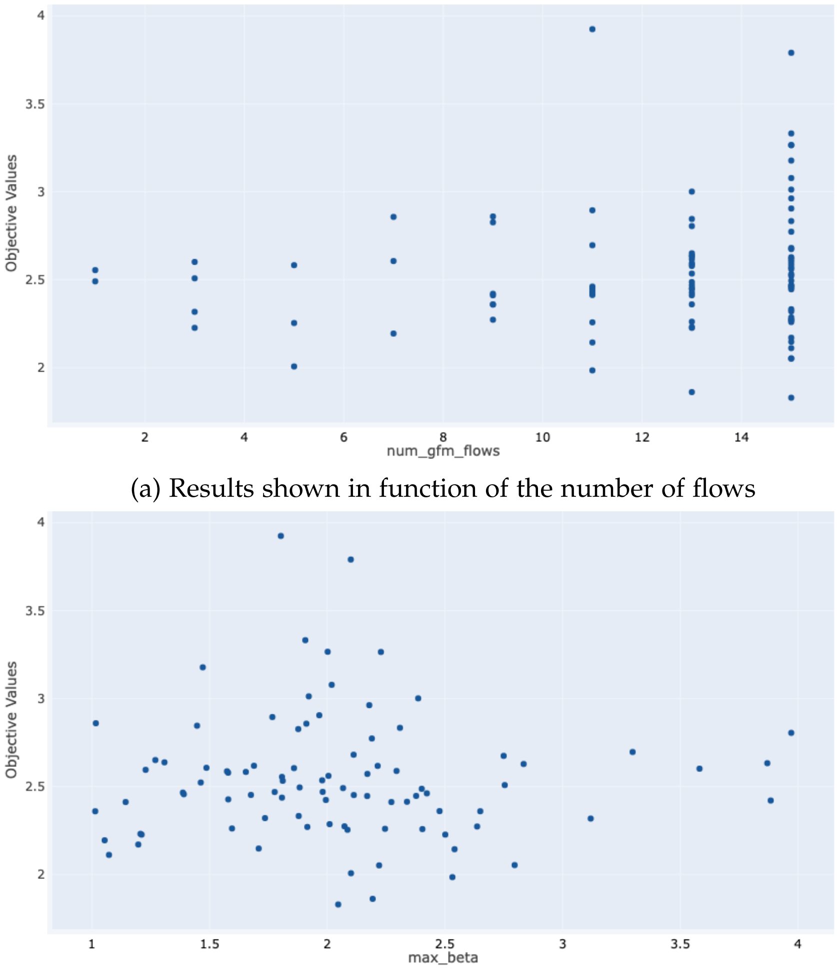


(a) Results shown in function of the number of coupling layers in each flow



(b) Results shown in function of the coupling layer dimension

Figure 6.2: Hyperoptimization run results for the Mopgfm in function of the number of coupling layers and coupling layer dimension. Every subfigure presents results in function of one parameter, with all other parameters varying.



(a) Results shown in function of the number of flows

(b) Results shown in function of $\beta$

Figure 6.3: Hyperoptimization run results for the Mopgfm method in function of the number of flows and $\beta$. Every subfigure presents results in function of one parameter, with all other parameters varying.

## 6.2 PolyMNIST

### 6.2.1 Evaluation of the Latent Representation

**Evaluation over epochs** Evaluating the separability of the latent representation (section 5.2.1) for models trained on 3 modalities, we find that the mofop, the mopgfm and the mopoe perform similarly, yielding on average a linear classification accuracy of 0.92, 0.91 and 0.92 respectively for all subsets after 500 training epochs (see fig. 6.4). The two methods that do not regularize the latent representation with the KL-divergence (iwmogfm, mogfm_amortized) perform worse than those that do, except for the moe and poe methods. The two latter methods have the worst performance overall.

**Evaluation across subset posterior approximations** Table 6.1 compares the classification accuracies of linear classifiers trained on each subset posterior. Overall, we see that the classification accuracy improves when more modalities make up the latent representation which shows that all methods are able to aggregate the modalities. In particular, we find that the iwmogfm method has the best performance when all modalities are given. Comparatively, the mopgfm is able to optimize the uni modal posteriors better than the mopoe

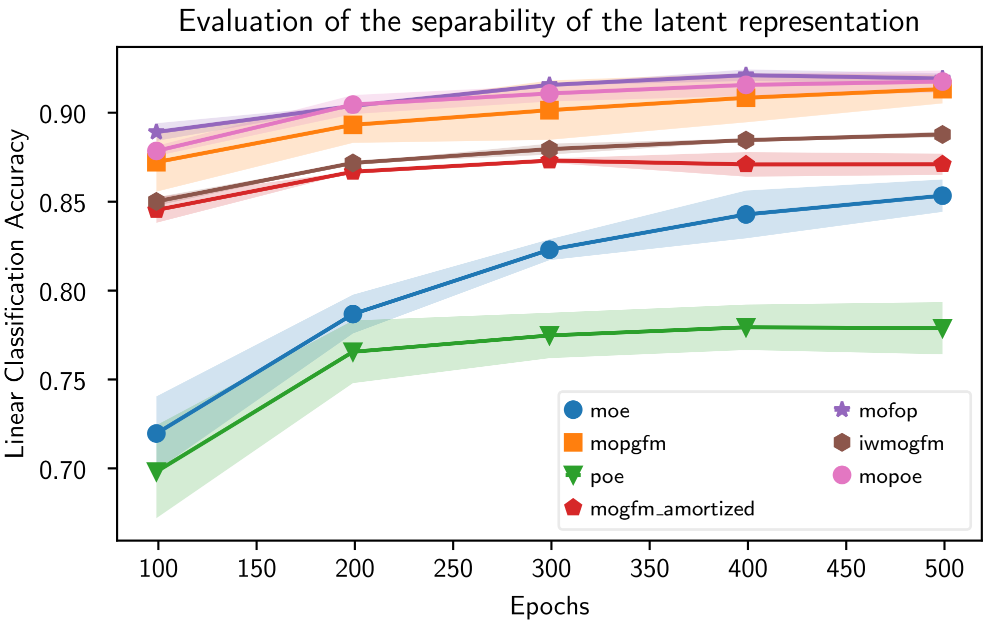


Figure 6.4: **Linear classification accuracy for different epochs over the test set, averaged over all subsets.** All methods were trained with 3 modalities.

| | m0 | m1 | m2 | m0_m1 | m0_m2 | m1_m2 | m0_m1_m2 |
|---|---|---|---|---|---|---|---|
| Method | | | | | | | |
| moe | 0.82±0.005 | 0.899±0.001 | 0.87±0.013 | 0.843±0.008 | 0.826±0.011 | 0.876±0.008 | 0.844±0.002 |
| mopgfm | **0.837**±0.004 | 0.928±0.007 | **0.924**±0.011 | 0.936±0.006 | 0.938±0.011 | 0.956±0.007 | 0.948±0.011 |
| poe | 0.224±0.016 | 0.893±0.062 | 0.723±0.098 | 0.861±0.059 | 0.69±0.1 | 0.957±0.005 | 0.941±0.012 |
| mogfm amortized | 0.615±0.007 | 0.84±0.008 | 0.824±0.005 | 0.905±0.003 | 0.904±0.007 | 0.96±0.003 | 0.96±0.003 |
| mofop | 0.783±0.004 | **0.936**±0.002 | 0.921±0.008 | 0.936±0.001 | 0.938±0.002 | **0.97**±0.002 | 0.953±0.0 |
| iwmogfm | 0.651±0.014 | 0.855±0.009 | 0.833±0.002 | 0.92±0.006 | 0.929±0.002 | 0.969±0.004 | **0.972**±0.003 |
| mopoe | 0.793±0.011 | 0.927±0.002 | 0.908±0.003 | **0.938**±0.005 | **0.942**±0.004 | 0.964±0.001 | 0.962±0.006 |

Table 6.1: Linear classification accuracy of all subset posterior approximations for the test set.

and the mofop, yielding an average accuracy of 0.896 compared to 0.876 and 0.88. Our results show that the $m0$ modality is the most difficult modality to learn from and as expected the poe struggles the most to optimize for it. It has the lowest accuracy on the subset containing only the $m0$ modality but compensates with the other modalities in the multi modal subsets. Similarly, both the iwmogfm and mogfm_amortized yield their lowest score on the $m0$ subset, while their performance improves significantly on the multi modal subsets.

**Scalability with the number of modalities** Figure 6.5 shows a comparison of how well each method scales with the number of modalities it is trained on, using the linear classification metric (section 5.2.1). Again, we see that the mofop, the mopoe and the mopgfm scale equally well with the number of modalities, the latter yielding a slightly better score for 1 modality.

### 6.2.2 Evaluation of the Generation Coherence

**Evaluation over epochs** Evaluating the generation coherence (section 5.2.2), we find that the mogfm_amortized and the iwmogfm perform the best overall, yielding an accuracy of $\approx 0.88$ after only 100 epochs (fig. 6.6). However, the performance of both methods does not improve after 100 epochs such that after 500 epochs, it almost matches that of the mopgfm and mofop. Overall, all methods making use of normalizing flow yield higher scores than the baseline methods.

**Comparison across missing modalities, reconstruction and random generation** For the generation coherence accuracy of missing modalities the mopgfm performs the best, followed by the mogfm_amortized and iwmogfm methods. For the reconstruction of modalities, both the mogfm_amortized and iwmogfm methods perform the best, followed by the mopgfm and mofop methods. For the generation of random samples, the moe provides a much higher coherence accuracy score than all other methods, implying that the moe learns a joint posterior that corresponds better to the prior than the other methods. Since the iwmogfm and mogfm_amortized were not trained

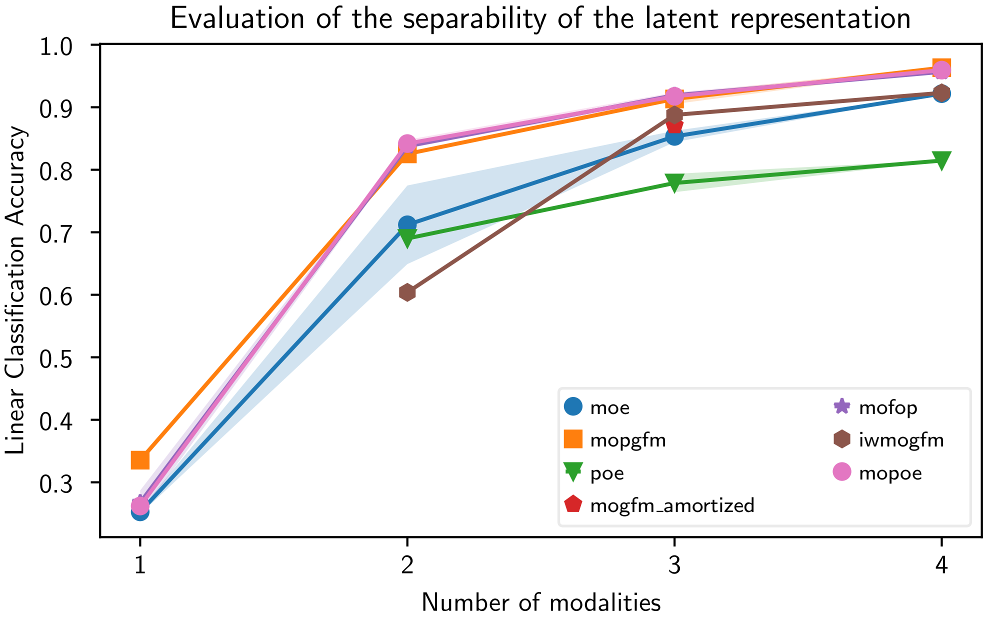


Figure 6.5: **Linear classification accuracy for models trained with different number of modalities, averaged over all subsets.** All methods were trained for 500 epochs.

with a regularization term in the objective that pushes their joint posterior to match the prior, their decoder networks do not recognize samples from the latter which explains the low accuracy for randomly generated images. Interestingly, the coherence accuracy of randomly generated samples with the mofop method is very low, suggesting that a higher regularization parameter $\beta$ might be needed.

**Scalability with the number of modalities** Overall, the mopgfm, the mopoe and the moe methods scale equally well with the number of modalities, the mopgfm yielding better performance than the mopoe, which itself performs better than the moe (fig. 6.7). For models trained on one modality, the coherence score is evaluated as self coherence only (section 5.2.2), which is an easier task than coherence across generated samples. This explains the slight dip in performance for all methods trained with 2 modalities.

### 6.2.3 Evaluation of the Generation Quality

Evaluating the generation quality (section 5.2.3), we find that overall the methods making use of the generalized $f$-mean perform the best. The mopgfm yields the best prd score for the generation of missing modalities, while

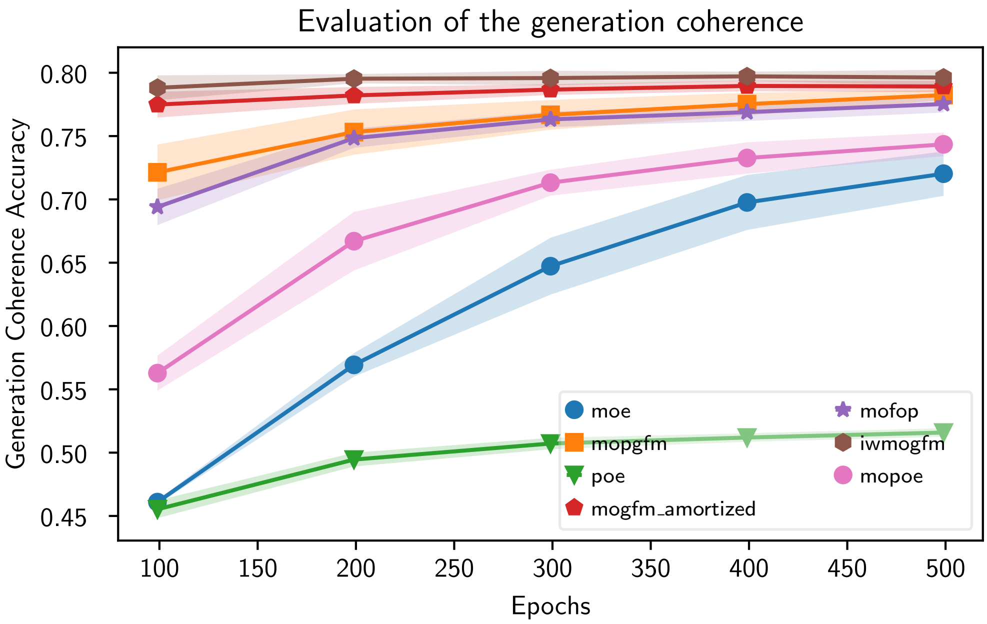


Figure 6.6: **Generation classification accuracy for different epochs over the test set, averaged over all combinations of input modalites and all output modalities.** All methods were trained with 3 modalities.

| Method | Missing Mod | Reconstruction | Random |
|---|---|---|---|
| moe | 0.732±0.011 | 0.742±0.013 | **0.303**±0.01 |
| mopgfm | **0.794**±0.003 | 0.834±0.005 | 0.156±0.003 |
| poe | 0.186±0.01 | 0.726±0.008 | 0.046±0.004 |
| mogfm amortized | 0.778±0.005 | 0.861±0.001 | 0.012±0.0 |
| mofop | 0.766±0.009 | 0.833±0.003 | 0.079±0.015 |
| iwmogfm | 0.774±0.009 | **0.869**±0.004 | 0.016±0.001 |
| mopoe | 0.727±0.008 | 0.796±0.005 | 0.194±0.017 |

Table 6.2: Coherence accuracy values evaluated for the generation of missing modalities, reconstruction of modalities and random generation on the Test set, for a model trained with 3 modalities. The coherence score for missing modalities is the average of all generation coherence accuracies for every subset of input modalities that does not contain the generated modality. Similarly, the coherence score for reconstruction is the average of all generation coherence accuracies for every subset of input modalities that *does* contain the generated modality.

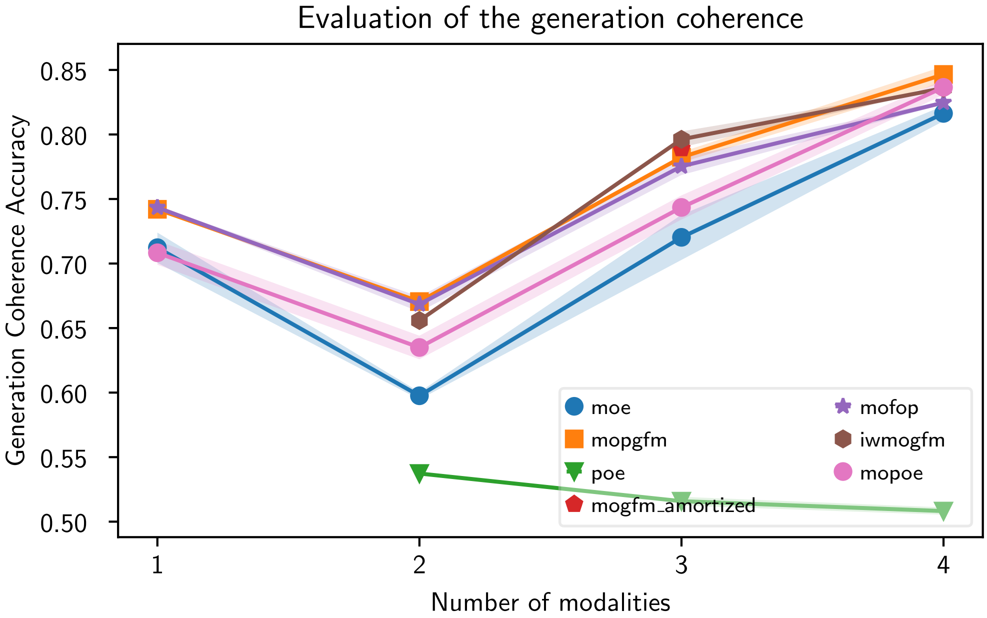


Figure 6.7: **Generation classification accuracy for models trained with different number of modalities.** The average over all classification accuracies is taken, across all possible combinations of input modalities and all output modalities, for three modalities from the PolyMNIST dataset.

both the mogfm_amortized and iwmogfm perform best on the reconstruction of modalities. Interestingly, the mopoe method provides prd scores with a higher variance than the other methods. The poe yields the best prd score for randomly generated samples followed by the mopoe, however a qualitative evaluation of randomly generated samples in fig. A.22 shows that while the mopoe generates the digits well, there is not much variance in the backgrounds. The modalitiy specific information (the background) of the randomly generated images from the mopoe actually seem to only correspond to an average of all pixels of the background image correspond to each modality. The same can be seen for the randomly generated images of the mopgfm and the moe. Overall the poe method captures best the modality specific information and provides the highest variance in the background of generated images.

Examples of generated samples for each method can be found in appendix A.

| | Missing Mod | Reconstruction | Random |
|---|---|---|---|
| Method | | | |
| moe | 0.094±0.007 | 0.2±0.002 | 0.125±0.005 |
| mopgfm | **0.283**±0.013 | 0.352±0.002 | 0.209±0.007 |
| poe | 0.263±0.032 | 0.329±0.035 | **0.261**±0.034 |
| mogfm amortized | 0.246±0.003 | **0.527**±0.005 | 0.0±0.0 |
| mofop | 0.128±0.01 | 0.333±0.009 | 0.157±0.002 |
| iwmogfm | 0.205±0.01 | 0.523±0.02 | 0.004±0.002 |
| mopoe | 0.135±0.048 | 0.334±0.044 | 0.222±0.025 |

Table 6.3: Area under the Precision and Recall curve of the PRD metric (Sajjadi et al., 2018) evaluated for the generation of missing modalities, reconstruction of modalities and random generation on the Test set, for a model trained with 3 modalities. The prd score for missing modalities is the average of all prd scores for every subset of input modalities that does not contain the generated modality. Similarly, the prd score for reconstruction is the average of all prd scores for every subset of input modalities that *does* contain the generated modality.

### 6.2.4 Comparison across different number of importance samples

Comparing how the mopoe, the mopgfm and the moe perform with different number of importance samples (K) (section 3.4) reveals that for the generation coherence (fig. 6.8) and the generation quality (fig. C.2), the performance of all three methods improves with a higher K. The evaluation on those two metrics also shows that the mopgfm performs better than the mopoe or the moe for any K. Interestingly, the generation coherence of the moe scales particularly well with a higher K, even surpassing the performance of the mopoe for $K = 3$ and $K = 5$. The improvement in the separability of the latent representation for a higher K is less clear (fig. C.1). A comparison of generated samples across different number of importance samples is shown in Appendix C.

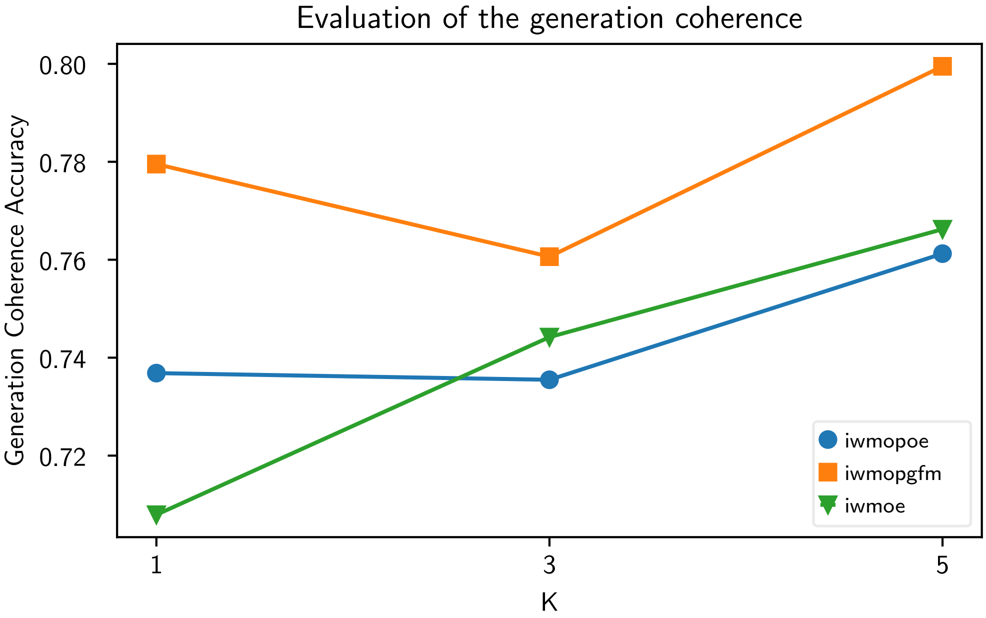
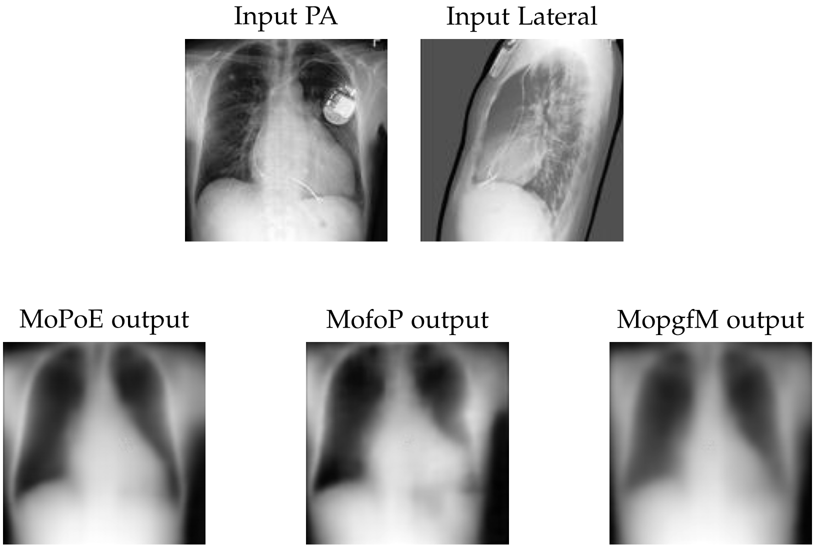


Figure 6.8: **Generation classification accuracy for models trained with different number of importance samples, evaluated on the PolyMNIST test set.** The average over all classification accuracies is taken, across all possible combinations of input modalities and all output modalities, for three modalities from the PolyMNIST dataset. All methods were trained with 3 modalities.

## 6.3 Mimic-CXR

A qualitative evaluation of generated samples from the MIMIC-CXR dataset reveals that the models are not able to capture smaller details in both the modality specific and the shared information. Figure 6.9 shows that the generated samples from the mopoe and mofop methods are extremely blurry and while approximately portraying the shape of the patient and its organs, smaller details like the ribs are lost. In fig. 6.9, the patient also has a support device, which is not represented in the generated samples. Figure B.14 shows that the quality of generated missing modalities is even worse, the generated Lateral modality being so blurred that it is hardly recognizable.



Input PA

Input Lateral

MoPoE output

MofoP output

MopgfM output

Figure 6.9: Comparison of conditionally generated PA samples from the mopoe and the mofop method. The generated samples are conditioned on images from a patient with a support device.

Chapter 7

# Conclusion & Discussion

We have implemented and tested new methods that provide a more flexible way to aggregate over multiple modalities in multi modal VAEs, using the generalized $f$-mean. Evaluating three metrics on the PolyMNIST dataset has shown that these methods improve results, especially for the coherence and the quality of the generated samples. This indicates that the generalized $f$-mean is able to better merge the information from each modality into a joint distribution than previous, fixed aggregation functions. However, a study of how well the methods scale with the number of modalities has shown that the methods utilizing normalizing flows scale less than those that do not. We hypothesise that this comes from the fact that each modalitiy is transformed with the same normalizing flow, such that with more modalities, the task of the flow to learn a meaningful mapping for each modality becomes increasingly difficult. We argue that this can be compensated with a higher amount of chained transformations, but which comes at a higher computational cost.

**MofoP & MopgfM** As introduced in section 4.3.1, the mofop builds on the mopoe by transforming each subset posterior approximation with a normalizing flow. While providing a more flexible joint posterior approximation, this does not make the aggregation over modalities more flexible, since the subset distributions are obtained with PoEs and the joint distribution with a MoE over subsets. We implemented and tested this method in comparison to our methods that utilize the generalized $f$-mean, to evaluate if the improved performance of those is due to a more flexible joint posterior distribution or a more flexible aggregation over modalities. The mopgfm provides a good comparison for this matter, since it utilizes the generalized $f$-mean, but uses a normal distributed posterior approximation. It thus has a flexible aggregation over modalities but does not have a more flexible joint posterior distribution. A comparison between the mofop and the mopgfm has shown that in general the mopgfm performs only slightly better than the mofop, indicating that both a more flexible joint posterior distribution and a more flexible aggregation function are able to improve results on the PolyMNIST dataset. This shows that transforming the subset posterior approximation of the mopgfm with normalizing flows to obtain a more flexible joint posterior distribution should further improve its results. Of course, this comes at the cost of increased computational cost and training time.

**mogfm_amortized & iwmogfm** The mogfm_amortized and the iwmofgm provide a way to obtain both, a more flexible aggregation function and a more flexible joint posterior approximation. The two methods make use of a modified objective to steer the joint posterior approximation towards a distribution that can be evaluated, but we have found this to be too unstable in practice. However, our results have shown that both methods are able to learn a good joint posterior distribution, even without the KL-divergence as regularization term in the objective (i.e. with $\beta = 0$). While this results in a very high generation coherence and quality, this also results in a less structured joint posterior distribution since both methods yield lower linear classification accuracies (section 5.2.1). In addition, since the joint posterior distribution of both methods cannot be evaluated explicitly, one cannot generate new data by sampling from it. Overall the mogfm_amortized and the iwmofgm provide very promising results and it would be interesting to evaluate in a more extensive study, if the weight of the regularization term in the objective can be adapted such that the learned posterior distribution of both methods matches a prior distribution.

A qualitative evaluation on the challenging MIMIC-CXR dataset shows that the methods are not able to extract meaningful information from the three provided modalitities. Independent of a more flexible joint posterior distribution and a more flexible aggregation over modalities, the generated samples are extremely blurry and fail to show details in both the modality specific and shared information. We argue that further adaptations to the training paradigm are needed to capture small details in real world datasets. Especially for medical images where the shared information between the modalities are pathologies that are sometimes hardly recognizable, even for human experts. In (Dorent et al., 2019), the authors show with their modified MVAE model, that aggregating over the modalities on multiple scales provides high quality results for the segmentation of brain tumours. This could be adapted for our more flexible aggregation function in future work.

Overall, we have shown that the generalized $f$-mean provides a great tool to improve the objective of multi modal VAEs. In future work, it would be interesting to evaluate theoretical properties of the more flexible aggregation function and how it impacts the tightness of the modified ELBO.

# Bibliography

Akiba, Takuya et al. (25th July 2019). “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *arXiv:1907.10902 [cs, stat]*. arXiv: 1907.10902. URL: http://arxiv.org/abs/1907.10902.

Baltrušaitis, T., C. Ahuja and L. Morency (Feb. 2019). “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 423–443. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2798607.

Berg, Rianne van den et al. (20th Feb. 2019). “Sylvester Normalizing Flows for Variational Inference”. In: *arXiv:1803.05649 [cs, stat]*. arXiv: 1803.05649. URL: http://arxiv.org/abs/1803.05649.

Bogachev, Vladimir I (2007). *Measure theory*. Vol. 1. Springer Science & Business Media.

Brubaker, Marcus (2020). *Introduction to Normalizing Flows (ECCV2020 Tutorial)*. https://youtu.be/u3vVyFVU_lI.

Burda, Yuri, Roger Grosse and Ruslan Salakhutdinov (7th Nov. 2016). “Importance Weighted Autoencoders”. In: *arXiv:1509.00519 [cs, stat]*. arXiv: 1509.00519. URL: http://arxiv.org/abs/1509.00519.

Burgess, Christopher P. et al. (10th Apr. 2018). “Understanding disentangling in $\beta$-VAE”. In: *arXiv:1804.03599 [cs, stat]*. version: 1. arXiv: 1804.03599. URL: http://arxiv.org/abs/1804.03599.

Chen, Ricky T. Q. et al. (23rd Apr. 2019). “Isolating Sources of Disentanglement in Variational Autoencoders”. In: *arXiv:1802.04942 [cs, stat]*. arXiv: 1802.04942. URL: http://arxiv.org/abs/1802.04942.

Daunhawer, Imant et al. (2021). “Self-supervised Disentanglement of Modality-Specific and Shared Factors Improves Multimodal Generative Models”. In: *Pattern Recognition*. Ed. by Zeynep Akata, Andreas Geiger and Torsten Sattler. Vol. 12544. Cham: Springer International Publishing, pp. 459–473. ISBN: 978-3-030-71277-8 978-3-030-71278-5. DOI: 10.1007/978-3-030-

71278-5_33. URL: https://link.springer.com/10.1007/978-3-030-71278-5_33.

Dependabot (2020). *Automated dependency updates.* https://dependabot.com.

Dinh, Laurent, David Krueger and Yoshua Bengio (10th Apr. 2015). "NICE: Non-linear Independent Components Estimation". In: *arXiv:1410.8516 [cs]*. arXiv: 1410.8516. URL: http://arxiv.org/abs/1410.8516.

Dinh, Laurent, Jascha Sohl-Dickstein and Samy Bengio (27th Feb. 2017). "Density estimation using Real NVP". In: *arXiv:1605.08803 [cs, stat]*. arXiv: 1605.08803. URL: http://arxiv.org/abs/1605.08803.

Dorent, Reuben et al. (25th July 2019). "Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation". In: *arXiv:1907.11150 [cs, eess]*. DOI: 10.1007/978-3-030-32245-8_9. arXiv: 1907.11150. URL: http://arxiv.org/abs/1907.11150.

Goodfellow, Ian J. et al. (10th June 2014). "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]*. arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Heusel, Martin et al. (26th June 2017). "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *arXiv:1706.08500 [cs, stat]*. version: 1. arXiv: 1706.08500. URL: http://arxiv.org/abs/1706.08500.

Higgins, Irina et al. (4th Nov. 2016). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hinton, Geoffrey E. (Aug. 2002). "Training products of experts by minimizing contrastive divergence". In: *Neural Computation* 14.8, pp. 1771–1800. ISSN: 0899-7667. DOI: 10.1162/089976602760128018.

Ioanas, Horea-Ioan and Markus Rudin (Aug. 2018). "Reproducible Self-Publishing for Python-Based Research". In: EuroSciPy. DOI: 10.6084/m9.figshare.7247339.v1. URL: https://figshare.com/articles/Reproducible_Self-Publishing_for_Python-Based_Research/7247339.

Johnson, Alistair EW et al. (2019). "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". In: *arXiv preprint arXiv:1901.07042*.

Karpathy, Andrej and Li Fei-Fei (2015). "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.html.

Kingma, Diederik P. and Jimmy Ba (29th Jan. 2017). "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]*. arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

Kingma, Diederik P. and Prafulla Dhariwal (10th July 2018). "Glow: Generative Flow with Invertible 1x1 Convolutions". In: *arXiv:1807.03039 [cs, stat]*. arXiv: 1807.03039. URL: http://arxiv.org/abs/1807.03039.

Kingma, Diederik P. and Max Welling (1st May 2014). "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]*. arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114.

Klug, Hendrik (2021a). *Data used for the Master Thesis "Multi Modal Generative Learning utilizing Normalizing Flows"*. DOI: 10.5281/ZENODO.5588294. URL: https://zenodo.org/record/5588294.

— (2021b). *The Reproducible Self Publishing toolkit for the Thesis "Multi Modal Normalizing Flows"*. https://github.com/Jimmy2027/MMNF_RepSeP.

Klug, Hendrik, Thomas M Sutter and Julia E Vogt (7th July 2021). "Multimodal Generative Learning on the MIMIC-CXR Database". In: MIDL 2021 Conference. URL: https://openreview.net/pdf?id=ZVqjoKVbYMl.

Ledig, Christian et al. (2017). "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Liu, Ming-Yu et al. (2019). "Few-shot unsupervised image-to-image translation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10551–10560.

Locatello, Francesco et al. (18th June 2019). "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *arXiv:1811.12359 [cs, stat]*. arXiv: 1811.12359. URL: http://arxiv.org/abs/1811.12359.

MMVAE_Hub (21st Oct. 2021). *Codebase for training multi modal VAEs on multiple datasets*. https://github.com/Jimmy2027/MMVAE_Hub.

Ngiam, Jiquan et al. (2011). "Multimodal Deep Learning". In: URL: https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf.

Nowozin, Sebastian (15th Feb. 2018). "Debiasing Evidence Approximations: On Importance-weighted Autoencoders and Jackknife Variational Inference". In: ICLR 2018 Conference. URL: https://www.microsoft.com/en-us/research/publication/debiasing-evidence-approximations-importance-weighted-autoencoders-jackknife-variational-inference/.

Pandey, Gaurav and Ambedkar Dukkipati (2017). "Variational methods for conditional multimodal deep learning". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 308–315.

Papamakarios, George et al. (5th Dec. 2019). "Normalizing Flows for Probabilistic Modeling and Inference". In: *arXiv:1912.02762 [cs, stat]*. arXiv: 1912.02762. URL: http://arxiv.org/abs/1912.02762.

Poore, Geoffrey M (2015). "PythonTeX: reproducible documents with LaTeX, Python, and more". In: *Computational Science & Discovery* 8.1, p. 014010.

doi: 10.1088/1749-4699/8/1/014010. url: https://iopscience.iop.org/article/10.1088/1749-4699/8/1/014010.

Pu, Yunchen et al. (2016). "Variational autoencoder for deep learning of images, labels and captions". In: *Advances in neural information processing systems* 29, pp. 2352–2360.

Rezende, Danilo Jimenez and Shakir Mohamed (14th June 2016). "Variational Inference with Normalizing Flows". In: *arXiv:1505.05770 [cs, stat]*. arXiv: 1505.05770. url: http://arxiv.org/abs/1505.05770.

Rezende, Danilo Jimenez, Shakir Mohamed and Daan Wierstra (18th June 2014). "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 1938-7228. PMLR, pp. 1278–1286. url: http://proceedings.mlr.press/v32/rezende14.html.

Sajjadi, Mehdi S. M. et al. (2018). "Assessing Generative Models via Precision and Recall". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Shi, Yuge et al. (8th Nov. 2019a). "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". In: *arXiv:1911.03393 [cs, stat]*. arXiv: 1911.03393. url: http://arxiv.org/abs/1911.03393.

Shi, Yuge et al. (2019b). "Variational mixture-of-experts autoencoders for multi-modal deep generative models". In: *Advances in Neural Information Processing Systems*, pp. 15718–15729.

Sutter, Thomas M., Imant Daunhawer and Julia E. Vogt (2nd Nov. 2020a). "Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence". In: *arXiv:2006.08242 [cs, stat]*. arXiv: 2006.08242. url: http://arxiv.org/abs/2006.08242.

Sutter, Thomas Marco, Imant Daunhawer and Julia E. Vogt (28th Sept. 2020b). "Generalized Multimodal ELBO". In: International Conference on Learning Representations. url: https://openreview.net/forum?id=5Y21V0RDBV.

Sutter, Thomas Marco, Imant Daunhawer and Julia E Vogt (2021). "Generalized Multimodal ELBO". In: *International Conference on Learning Representations*.

Suzuki, Masahiro, Kotaro Nakayama and Yutaka Matsuo (2016). "Joint multimodal learning with deep generative models". In: *arXiv preprint arXiv:1611.01891*.

Travis, CI (2011). *Hosted continuous integration service used to build and test software projects hosted on GitHub and Bitbucket*. https://www.travis-ci.com.

Tsai, Yao-Hung Hubert et al. (27th Sept. 2018a). "Learning Factorized Multimodal Representations". In: International Conference on Learning Representations. url: https://openreview.net/forum?id=rygqqsA9KX.

— (2018b). "Learning factorized multimodal representations". In: *arXiv preprint arXiv:1806.06176*.

Wikipedia (21st Oct. 2021). *Sum of normally distributed random variables*. https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables.

Wu, Mike and Noah Goodman (12th Nov. 2018). "Multimodal Generative Models for Scalable Weakly-Supervised Learning". In: *arXiv:1802.05335 [cs, stat]*. arXiv: 1802.05335. URL: http://arxiv.org/abs/1802.05335.

— (10th Dec. 2019). "Multimodal Generative Models for Compositional Representation Learning". In: *arXiv:1912.05075 [cs, stat]*. arXiv: 1912.05075. URL: http://arxiv.org/abs/1912.05075.

Appendix A

# Qualitative comparison of generated PolyMNIST samples

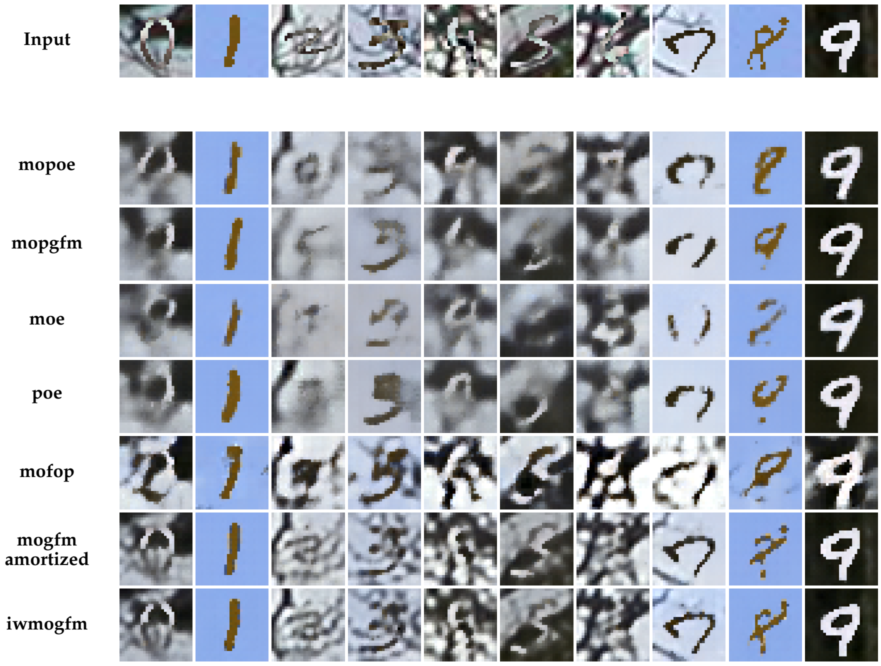


Figure A.1: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m0.

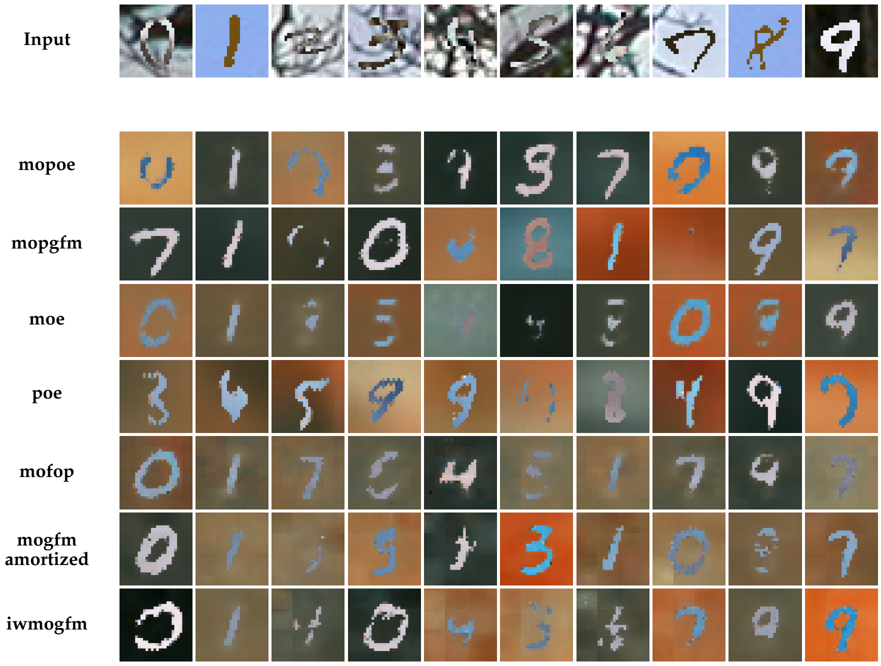


Figure A.2: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m1.

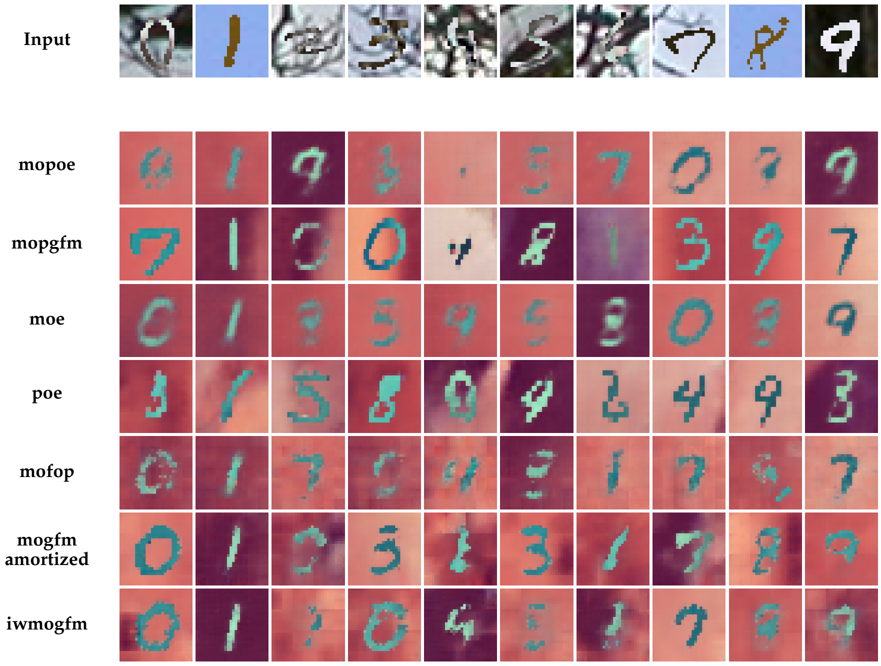


Figure A.3: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m2.

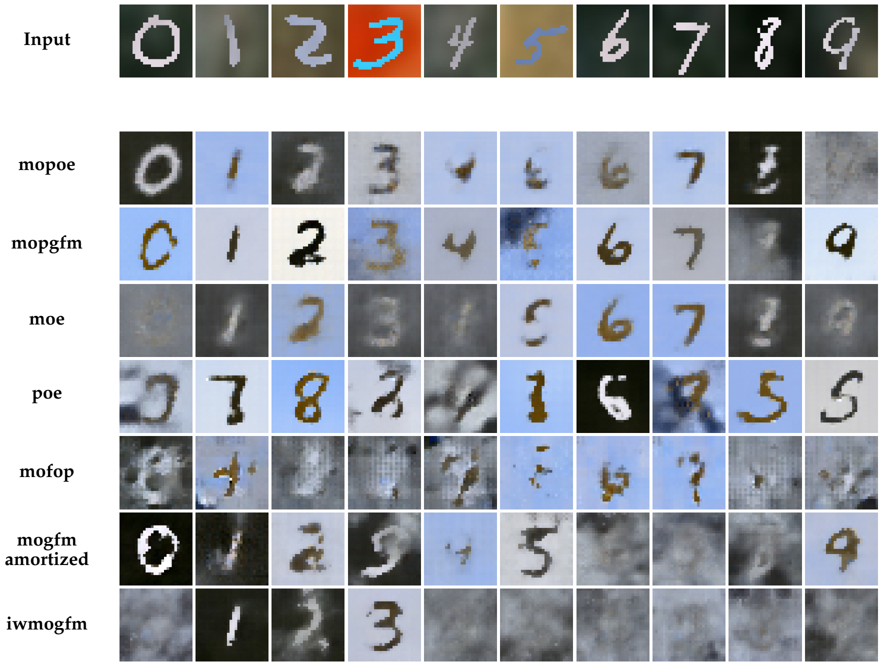


Figure A.4: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m0.

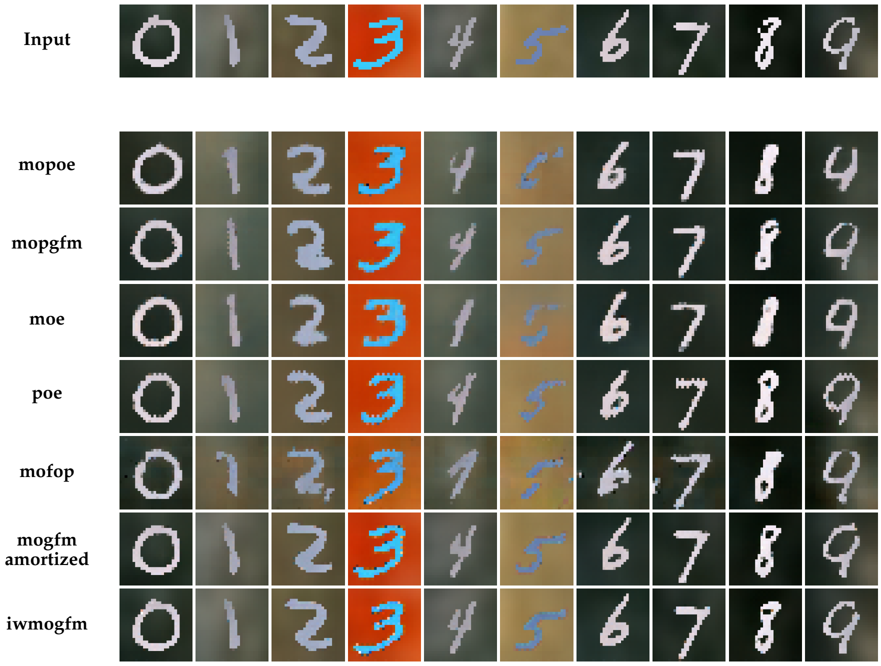


Figure A.5: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m1.

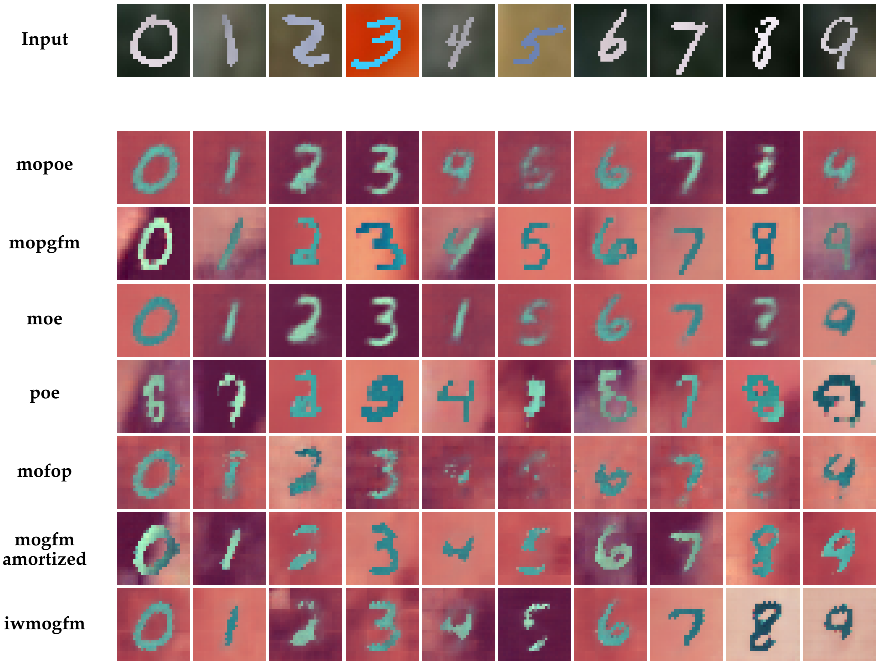


Figure A.6: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m2.

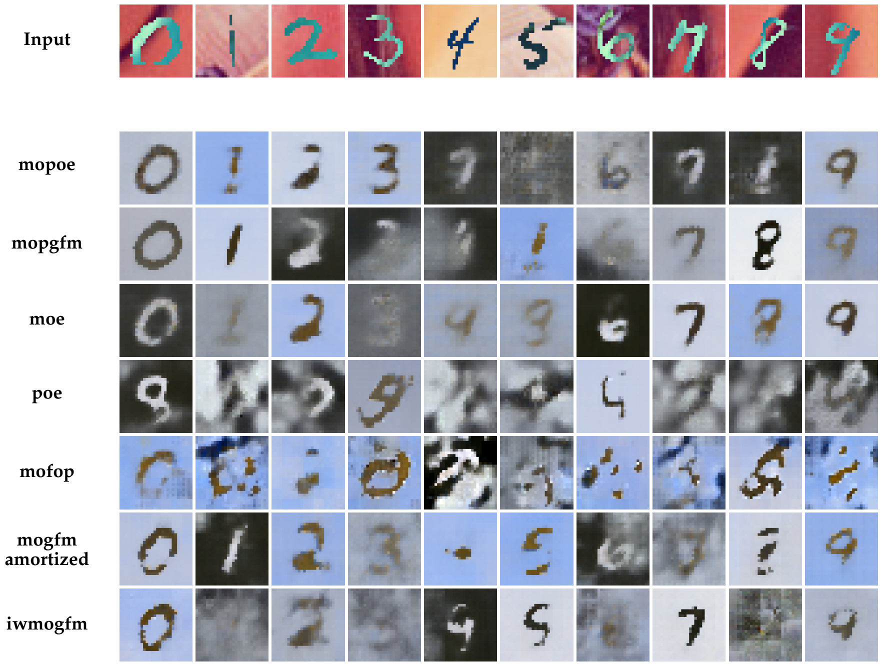


Figure A.7: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m0.

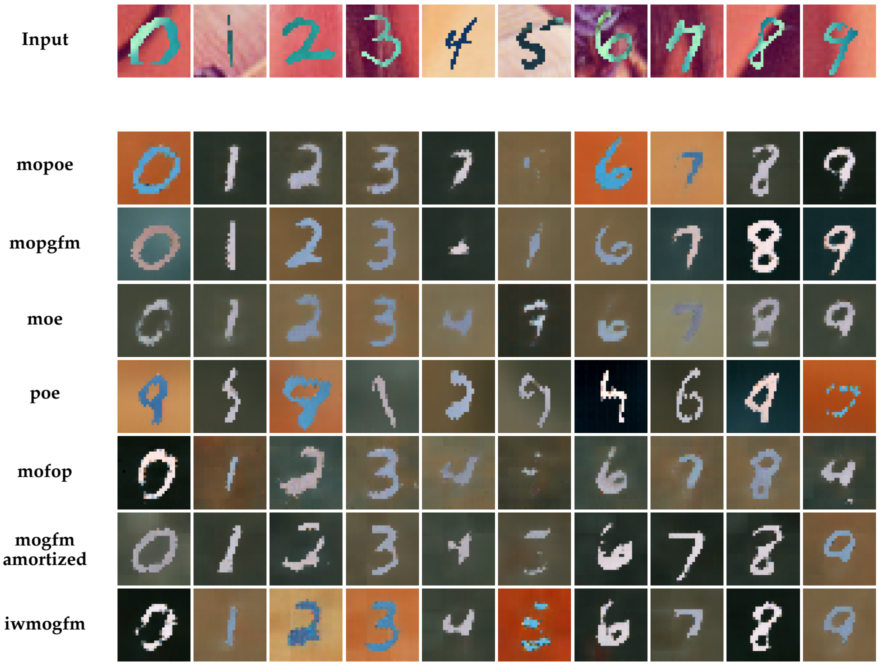


Figure A.8: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m1.

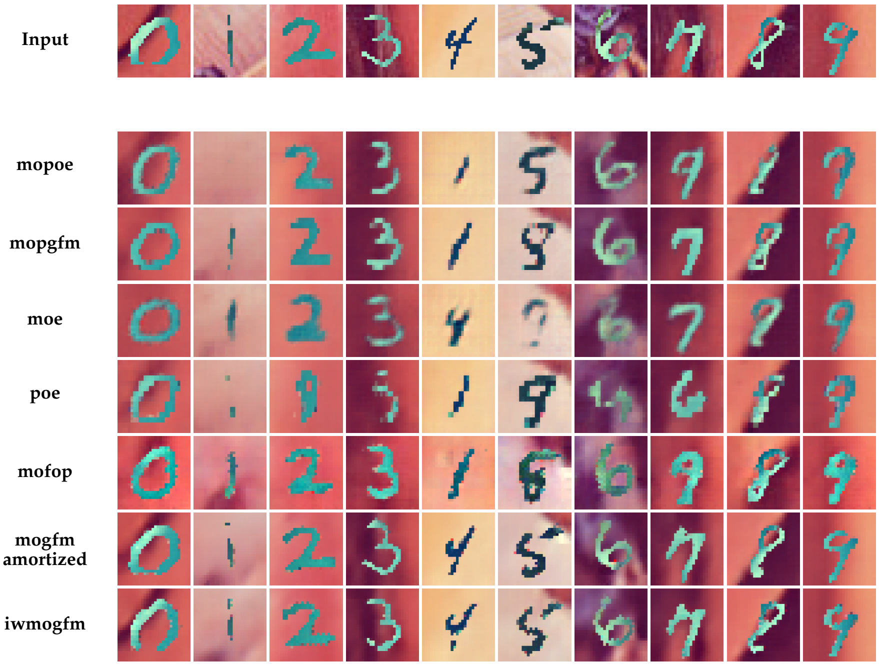


Figure A.9: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m2.

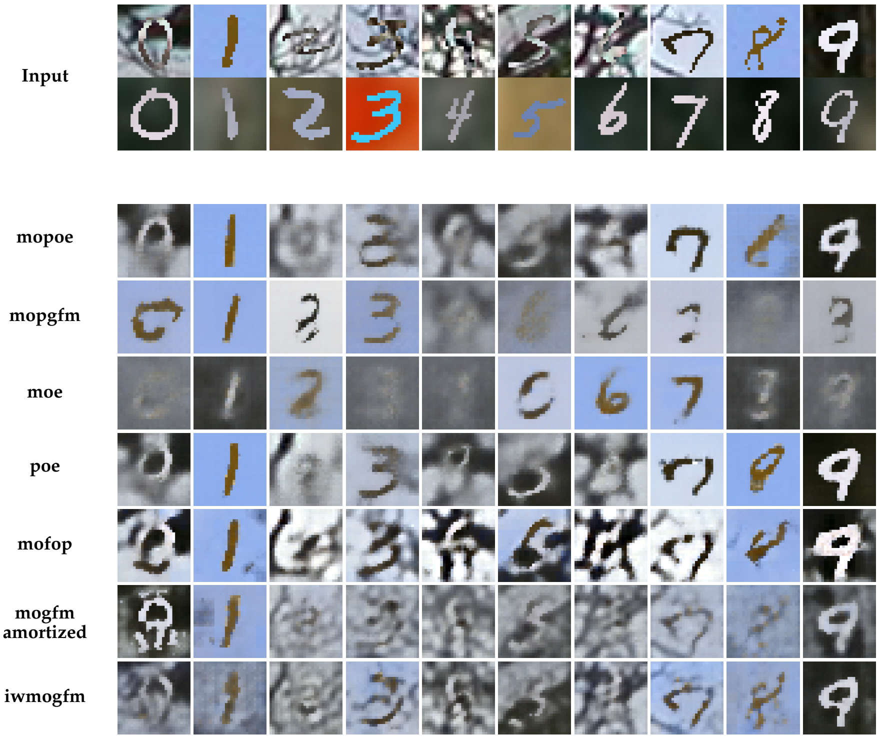


Figure A.10: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m0.

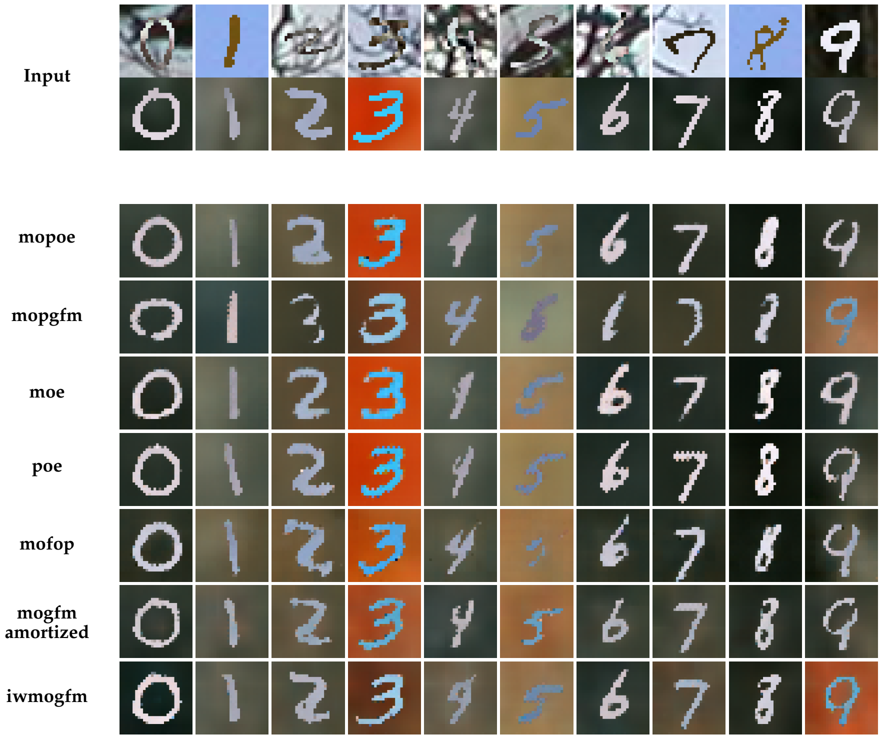


Figure A.11: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m1.

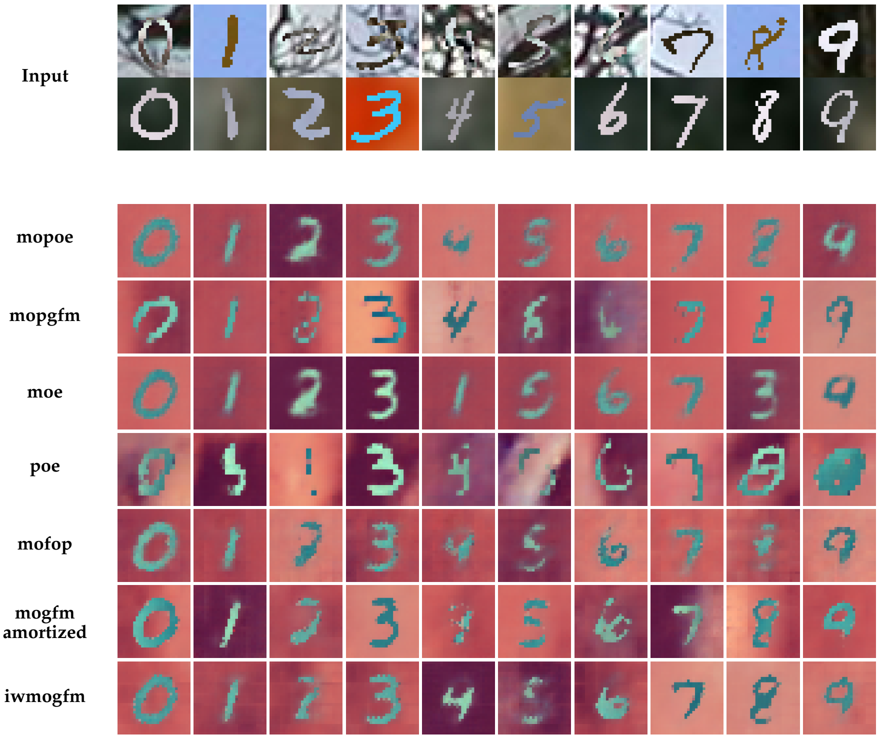


Figure A.12: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m2.

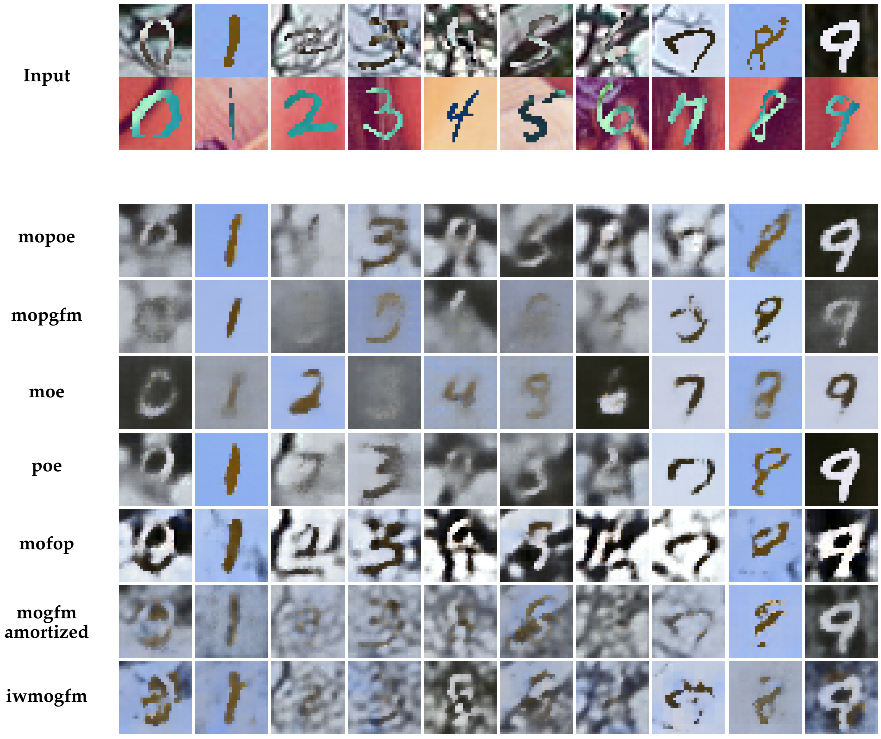


Figure A.13: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m0.

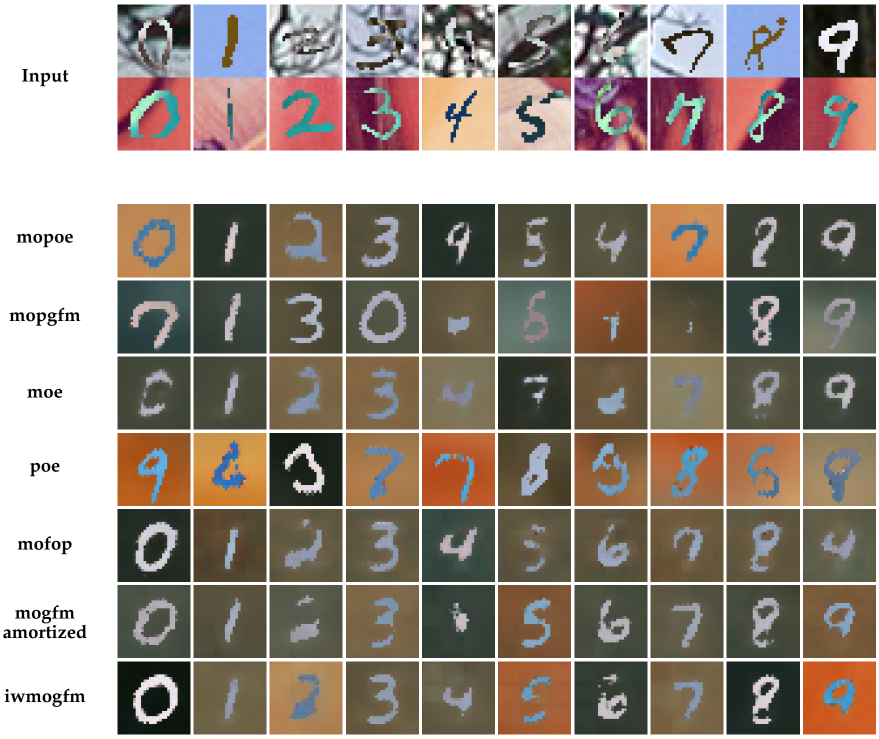


Figure A.14: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m1.

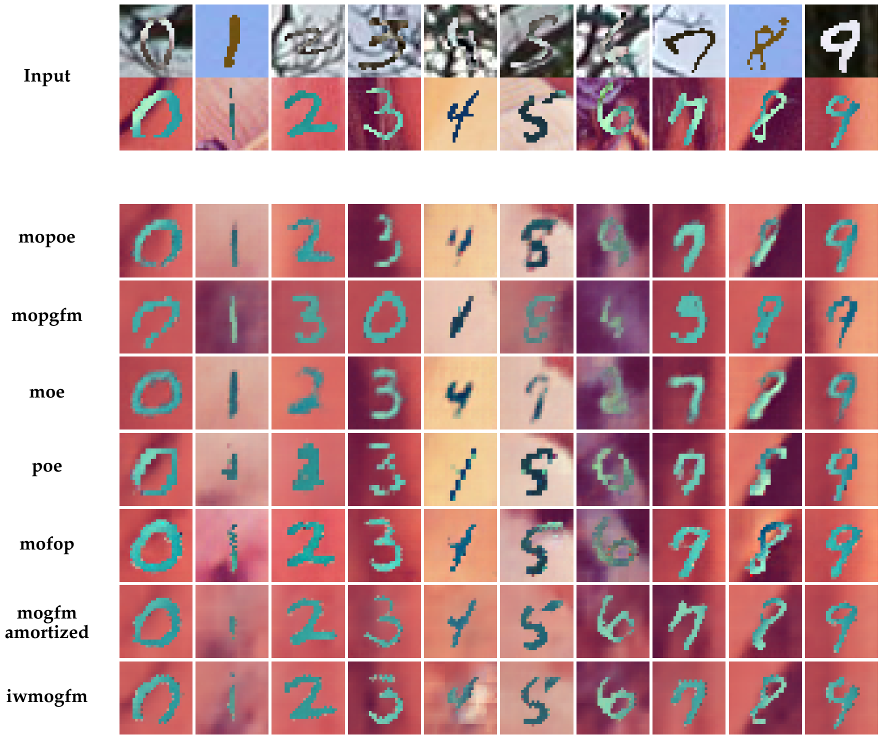


Figure A.15: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m2.

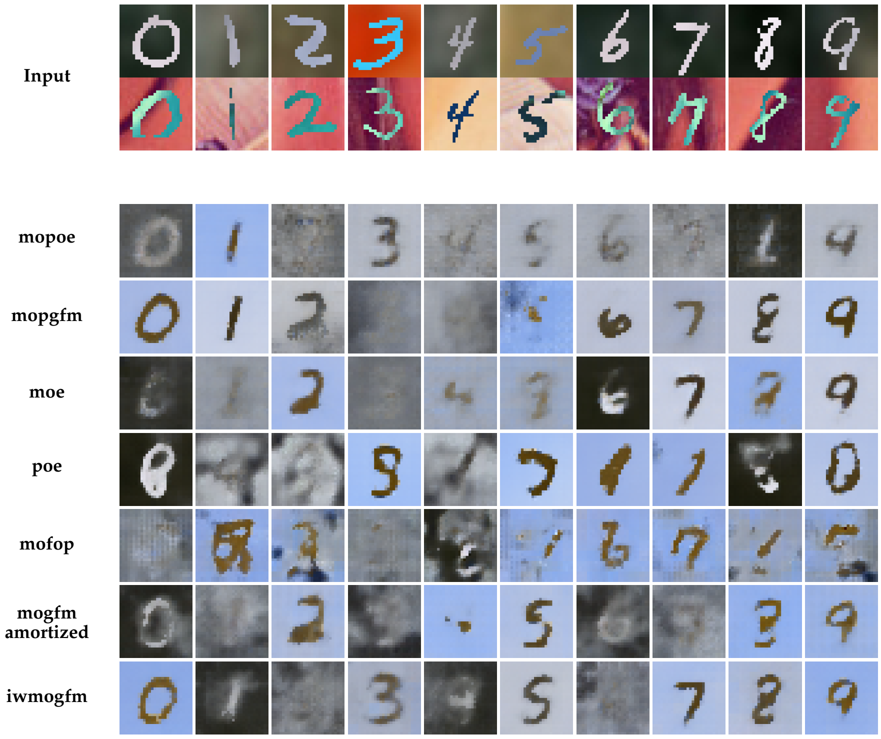


Figure A.16: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m0.

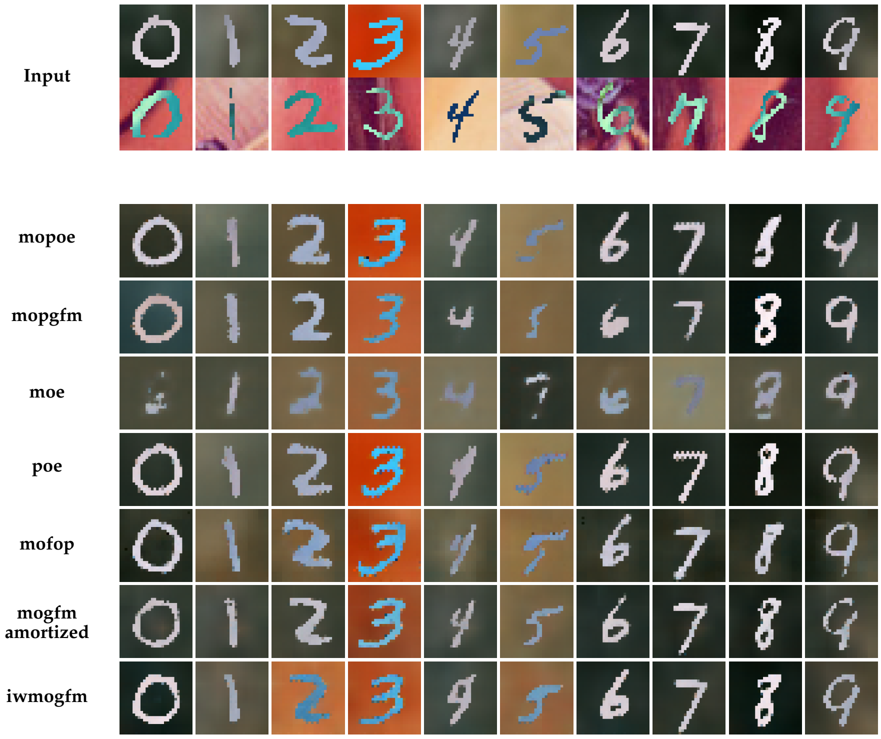


Figure A.17: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m1.

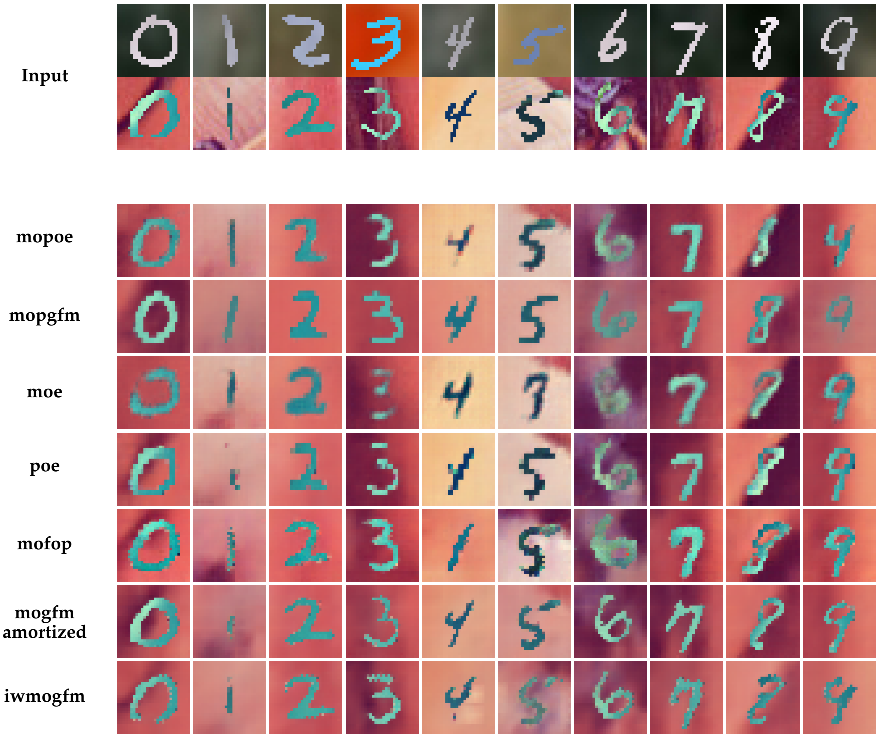


Figure A.18: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m2.

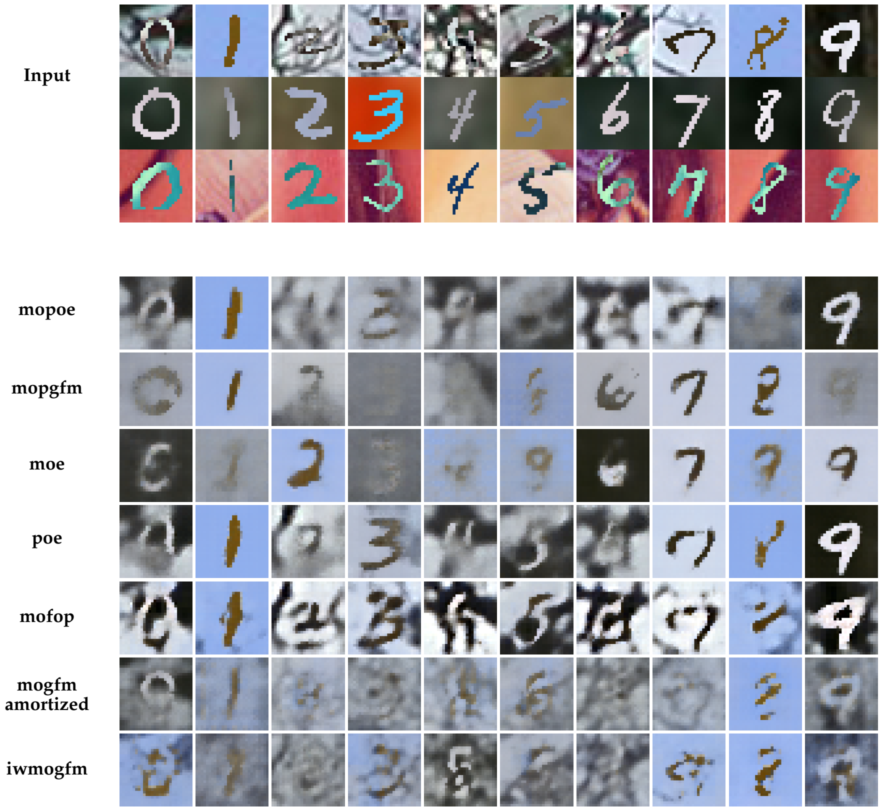


Figure A.19: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m0.

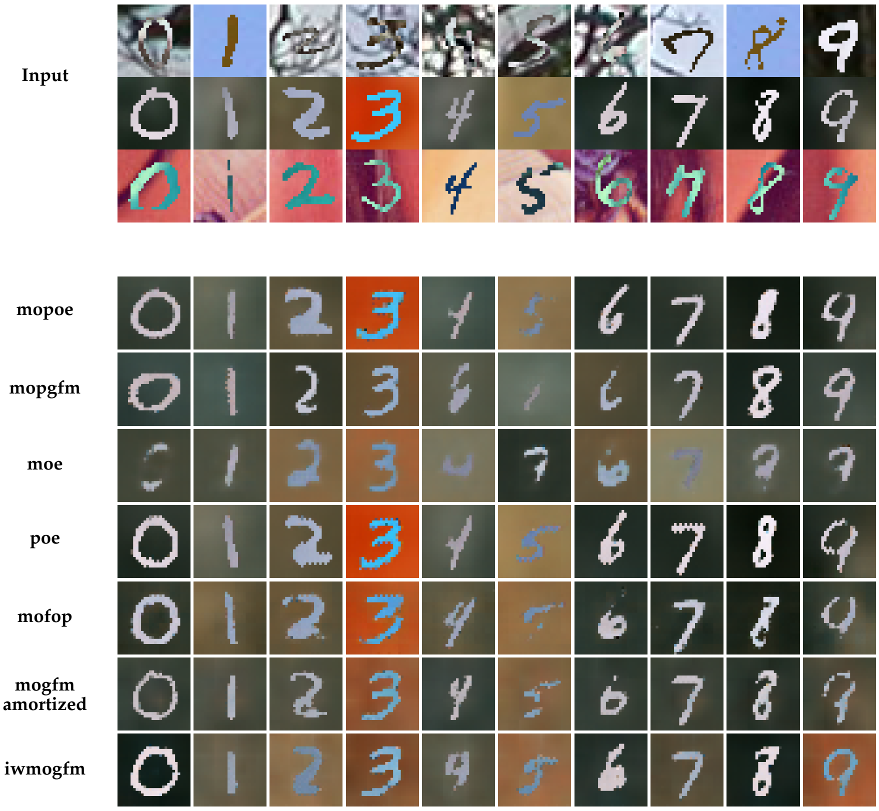


Figure A.20: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m1.

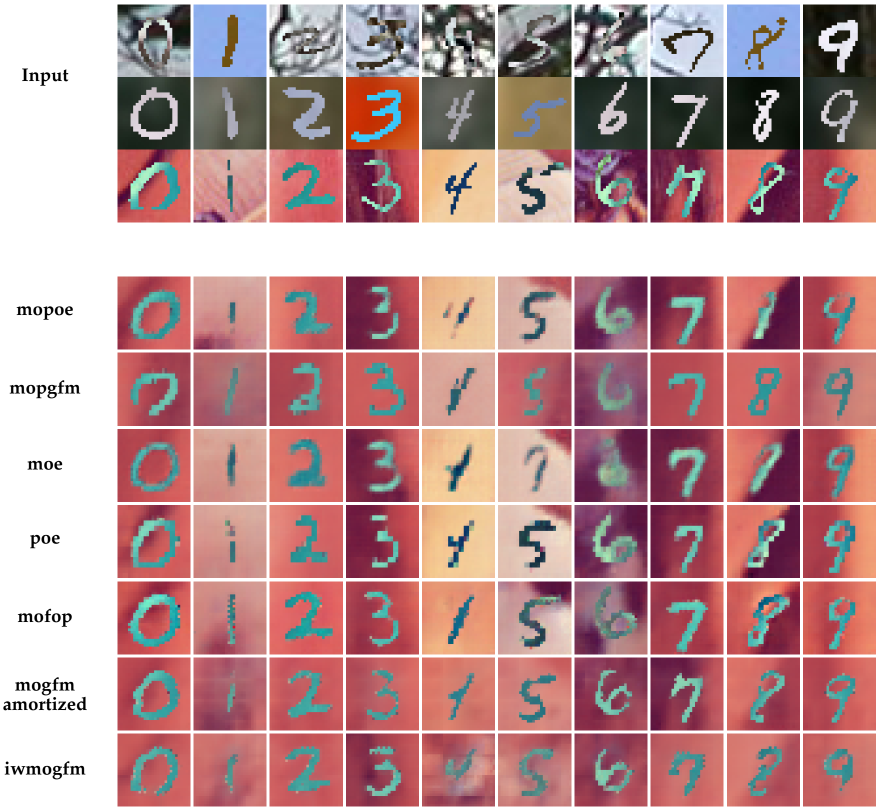


Figure A.21: Generated examples, conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m2.

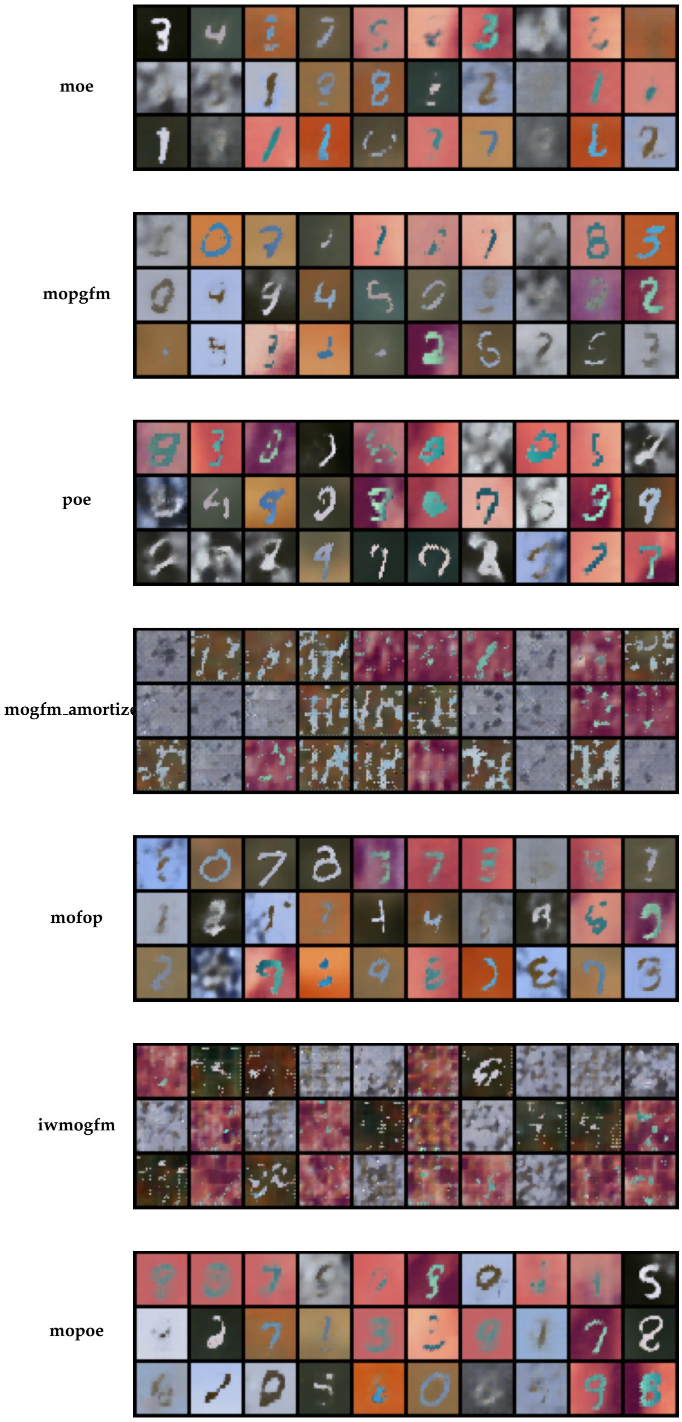


Figure A.22: **Comparison of randomly generated samples between methods.** The samples are generated by sampling from the prior and decoding them with a randomly selected decoder from the modalities $m_0$, $m_1$, $m_2$.

Appendix B

# Qualitative comparison of generated Mimic-CXR samples

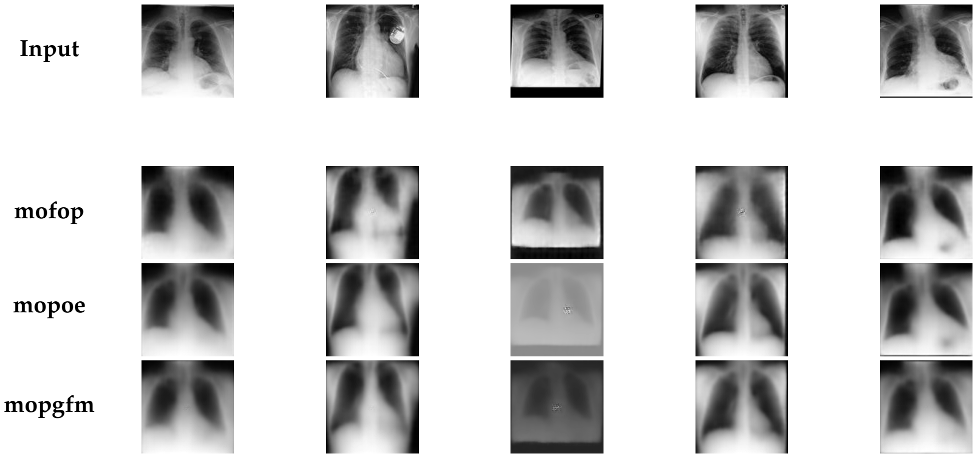


Figure B.1: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is PA and the generated modality is PA.

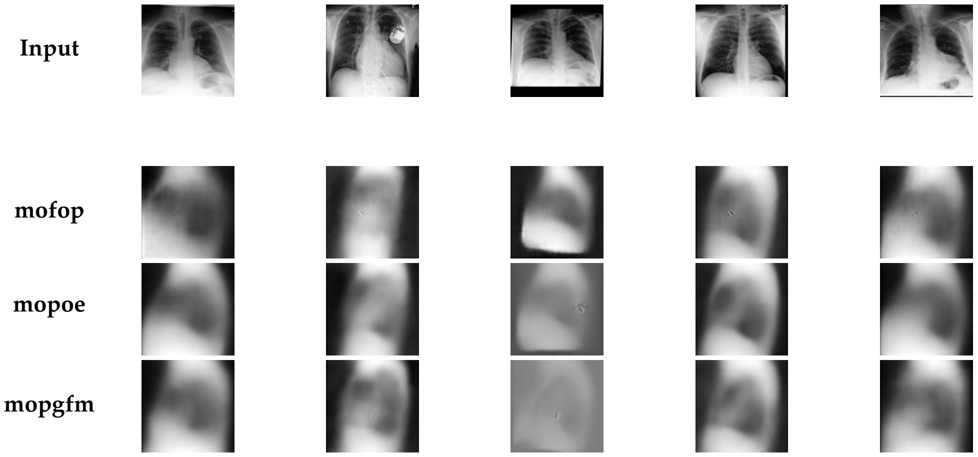


Figure B.2: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is PA and the generated modality is Lateral.

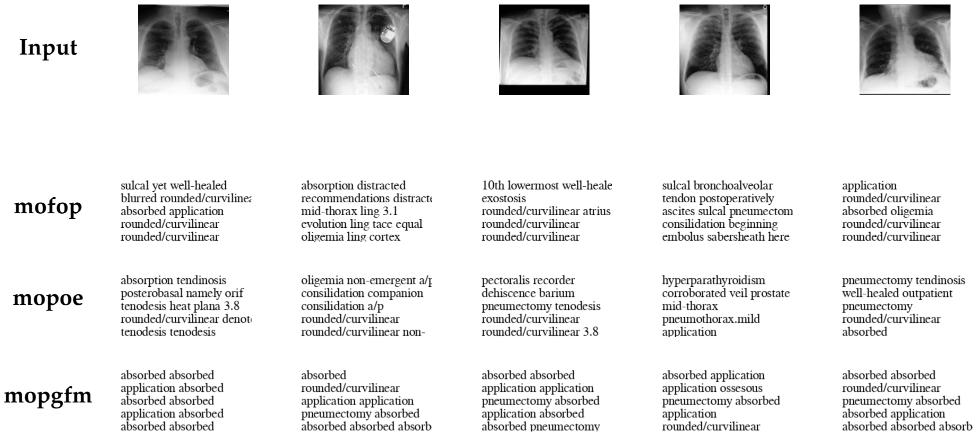

Input
mofop
sulcal yet well-healed blurred rounded/curvilinea absorbed application rounded/curvilinear rounded/curvilinear
absorption distracted recommendations distracte mid-thorax ling 3.1 evolution ling tace equal oligemia ling cortex
10th lowermost well-heale exostosis rounded/curvilinear atrius rounded/curvilinear rounded/curvilinear
sulcal bronchoalveolar tendon postoperatively ascites sulcal pneumectom consilidation beginning embolus sabersheath here
application rounded/curvilinear absorbed oligemia rounded/curvilinear rounded/curvilinear
mopoe
absorption tendinosis posterobasal namely orif tenodesis heat plana 3.8 rounded/curvilinear denot tenodesis tenodesis
oligemia non-emergent a/p consilidation companion consilidation a/p rounded/curvilinear rounded/curvilinear non-
pectoralis recorder dehiscence barium pneumectomy tenodesis rounded/curvilinear rounded/curvilinear 3.8
hyperparathyroidism corroborated veil prostate mid-thorax pneumothorax.mild application
pneumectomy tendinosis well-healed outpatient pneumectomy rounded/curvilinear absorbed
mopgfm
absorbed absorbed application absorbed absorbed absorbed application absorbed absorbed absorbed
absorbed rounded/curvilinear application application pneumectomy absorbed absorbed absorbed absorb
absorbed absorbed application application pneumectomy absorbed application absorbed absorbed pneumectomy
absorbed application application ossesous pneumectomy absorbed application rounded/curvilinear
absorbed absorbed rounded/curvilinear pneumectomy absorbed absorbed application absorbed absorbed absorb

Figure B.3: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is PA and the generated modality is text.

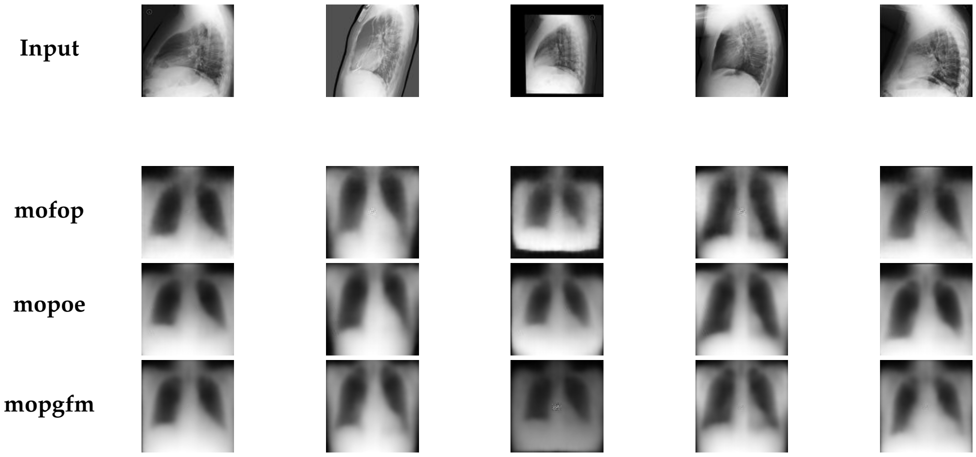


Figure B.4: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is Lateral and the generated modality is PA.

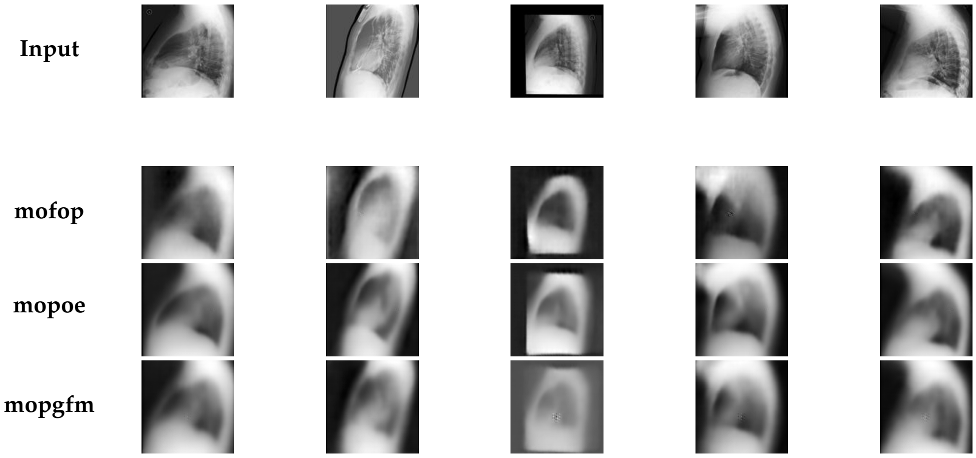
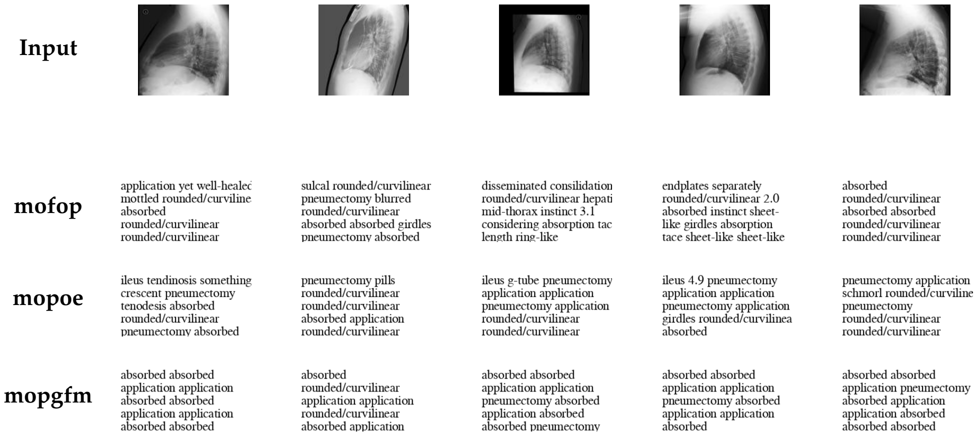
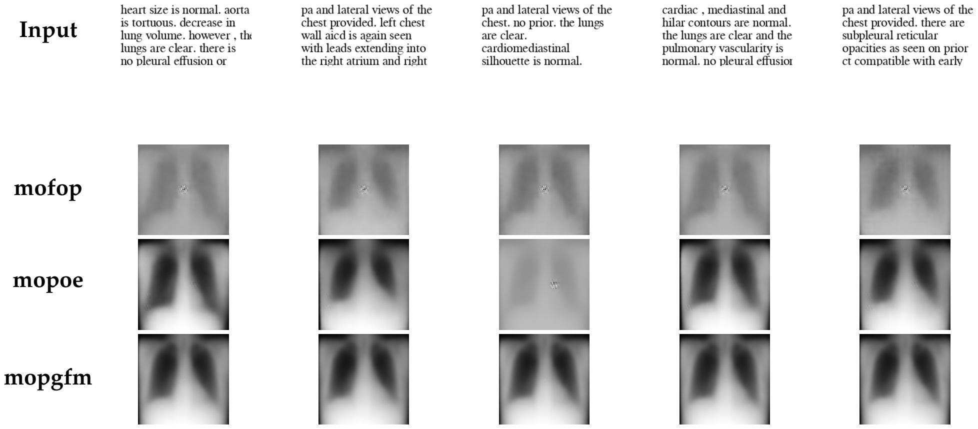


Figure B.5: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is Lateral and the generated modality is Lateral.



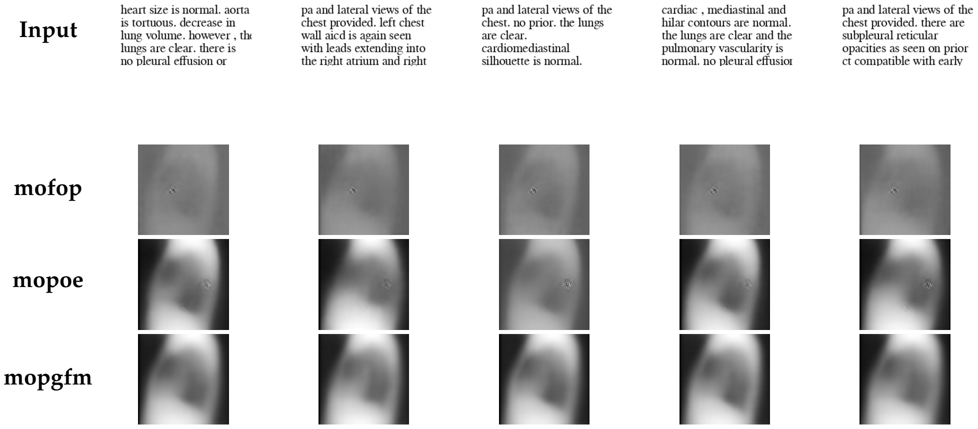
| | | | | | |
|---|---|---|---|---|---|
| mofop | application yet well-healed mottled rounded/curviline absorbed rounded/curvilinear rounded/curvilinear | sulcal rounded/curvilinear pneumectomy blurred rounded/curvilinear absorbed absorbed girdles pneumectomy absorbed | disseminated consilidation rounded/curvilinear hepati mid-thorax instinct 3.1 considering absorption tac length ring-like | endplates separately rounded/curvilinear 2.0 absorbed instinct sheet-like girdles absorption tace sheet-like sheet-like | absorbed rounded/curvilinear absorbed absorbed rounded/curvilinear rounded/curvilinear |
| mopoe | ileus tendinosis something crescent pneumectomy tenodesis absorbed rounded/curvilinear pneumectomy absorbed | pneumectomy pills rounded/curvilinear rounded/curvilinear absorbed application rounded/curvilinear | ileus g-tube pneumectomy application application pneumectomy application rounded/curvilinear rounded/curvilinear | ileus 4.9 pneumectomy application application pneumectomy application girdles rounded/curvilinea absorbed | pneumectomy application schmorl rounded/curviline pneumectomy rounded/curvilinear rounded/curvilinear |
| mopgfm | absorbed absorbed application application absorbed absorbed application application absorbed absorbed | absorbed rounded/curvilinear application application rounded/curvilinear absorbed application | absorbed absorbed application application pneumectomy absorbed application absorbed absorbed pneumectomy | absorbed absorbed application application pneumectomy absorbed application application absorbed | absorbed absorbed application pneumectomy absorbed application application absorbed absorbed absorbed |

Figure B.6: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is Lateral and the generated modality is text.



| | | | | | |
|---|---|---|---|---|---|
| Input | heart size is normal. aorta is tortuous. decrease in lung volume. however , the lungs are clear. there is no pleural effusion or | pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right | pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal. | cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusion | pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early |

Figure B.7: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is text and the generated modality is PA.



| | | | | | |
|---|---|---|---|---|---|
| **Input** | heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or | pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right | pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal. | cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusior | pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early |

Figure B.8: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is text and the generated modality is Lateral.

| | | | | | |
|---|---|---|---|---|---|
| **Input** | heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or | pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right | pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal. | cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusior | pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early |
| **mofop** | ileus separately rounded/curvilinear thickenings absorbed instinct sheet-like determined absorption tac | sulcal lowermost enlarge blurred ascites hemoptysis spleen sabersheath meduastinal alone sabersheath here counters | ileus separately rounded/curvilinear thickenings absorbed instinct sheet-like dystrophic absorption tace | medical separately posterobasal superior-mos posture separately sheet-like determined gallstones t2 sheet-like girdles | sulcal hip tendon mottled rounded/curvilinear neo loculared rounded/curvilinear mottle rounded/curvilinear |
| **mopoe** | rounded/curvilinear 4.9 loculared diuresis application pneumectomy application unknown rounded/curvilinear | pericardium tendinosis repositioning mid-thorax pneumectomy t3 closed oligemia absorption alone sabersheath final final | pericardium tendinosis laminectomies namely stat closed intracavitary among rounded/curvilinear beginning original | pericardium etiologies posterobasal cut subjacent bronchoalveolar efusions plana efusions hooks preservation | pericardium tendinosis repositioning mid-thorax pneumectomy t3 loculared oligemia rounded/curvilinear |
| **mopgfm** | absorbed absorbed application application pneumectomy absorbed application application application application | absorbed application application absorbed absorbed absorbed absorb absorbed absorbed absorb absorbed absorbed | absorbed absorbed application application pneumectomy application application application application application | absorbed absorbed application application pneumectomy absorbed application application absorbed | absorbed absorbed application application pneumectomy absorbed application absorbed absorbed absorbed |

Figure B.9: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modality is text and the generated modality is text.

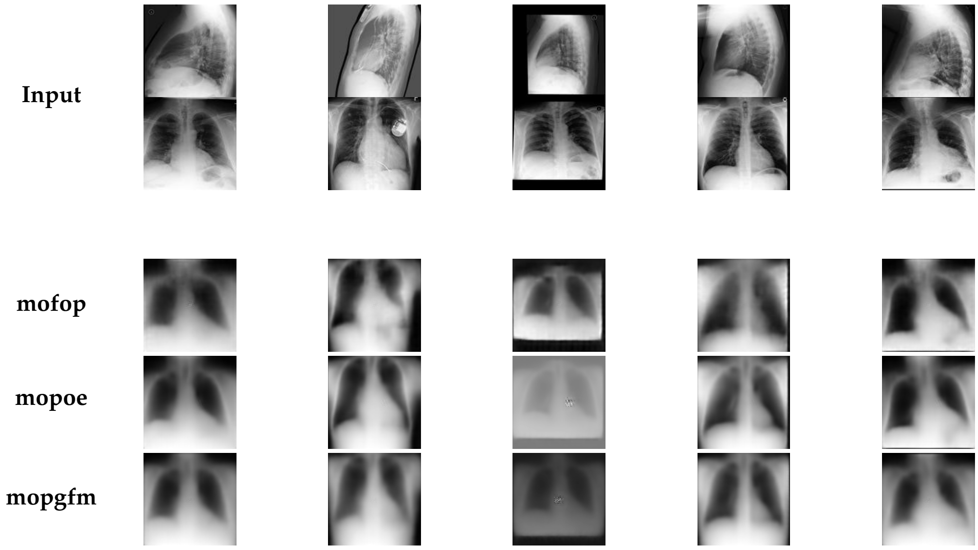


Figure B.10: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA and the generated modality is PA.

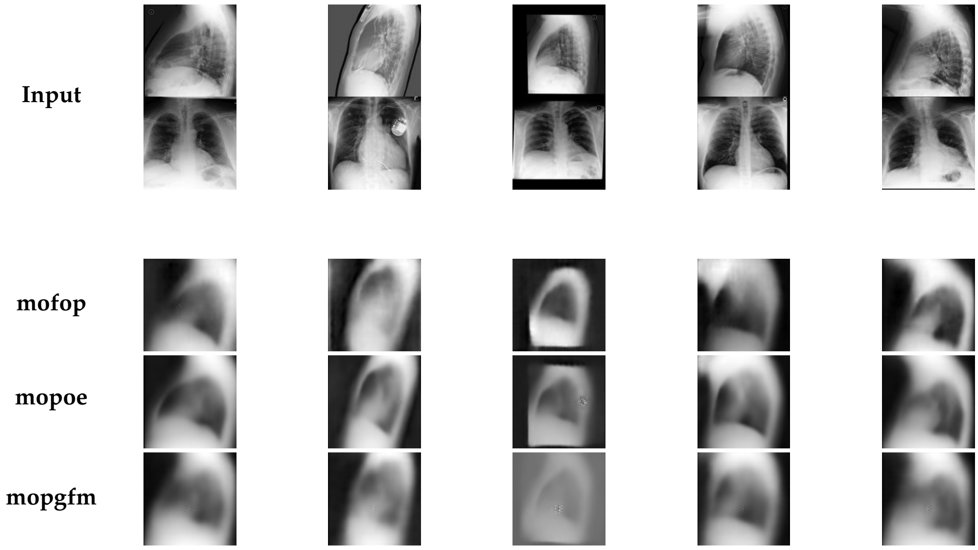


Figure B.11: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA and the generated modality is Lateral.

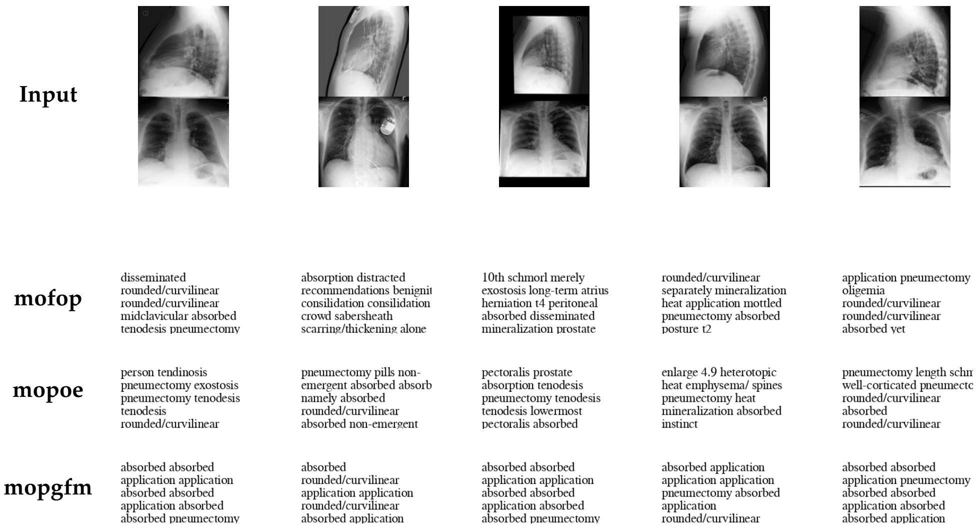

Input
mofop
disseminated rounded/curvilinear rounded/curvilinear midclavicular absorbed tenodesis pneumectomy
absorption distracted recommendations benignit consilidation consilidation crowd sabersheath scarring/thickening alone
10th schmorl merely exostosis long-term atrius herniation t4 peritoneal absorbed disseminated mineralization prostate
rounded/curvilinear separately mineralization heat application mottled pneumectomy absorbed posture t2
application pneumectomy oligemia rounded/curvilinear rounded/curvilinear absorbed yet
mopoe
person tendinosis pneumectomy exostosis pneumectomy tenodesis tenodesis rounded/curvilinear
pneumectomy pills non-emergent absorbed absorb namely absorbed rounded/curvilinear absorbed non-emergent
pectoralis prostate absorption tenodesis pneumectomy tenodesis tenodesis lowermost pectoralis absorbed
enlarge 4.9 heterotopic heat emphysema/ spines pneumectomy heat mineralization absorbed instinct
pneumectomy length schn well-corticated pneumectc rounded/curvilinear absorbed rounded/curvilinear
mopgfm
absorbed absorbed application application absorbed absorbed application absorbed absorbed pneumectomy
absorbed rounded/curvilinear application application rounded/curvilinear absorbed application
absorbed absorbed application application absorbed absorbed application absorbed absorbed pneumectomy
absorbed application application application pneumectomy absorbed application rounded/curvilinear
absorbed absorbed application pneumectomy absorbed absorbed application absorbed absorbed application

Figure B.12: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA and the generated modality is text.

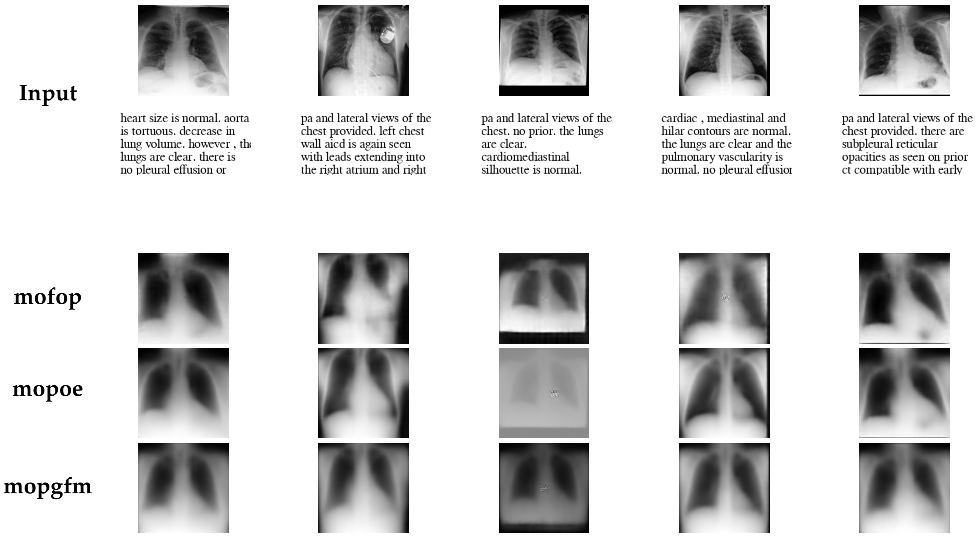
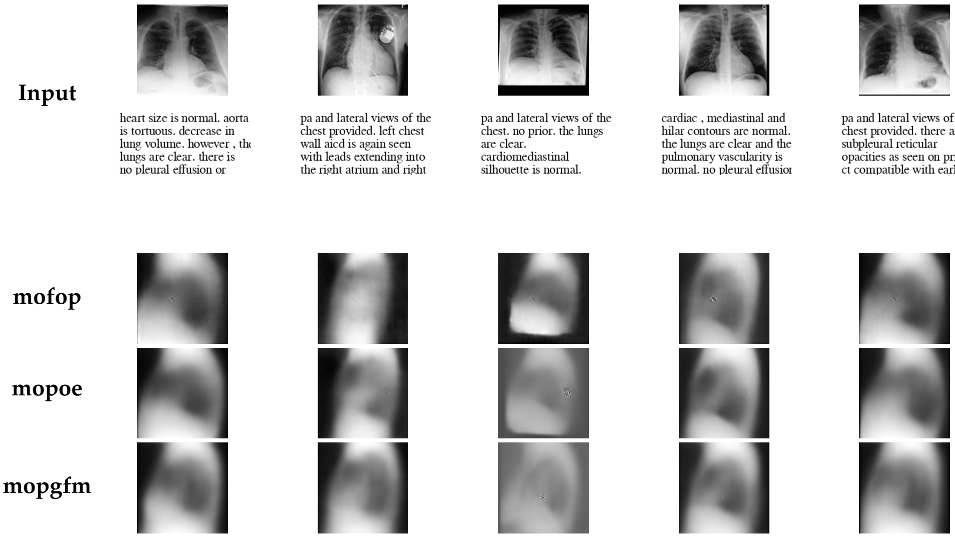
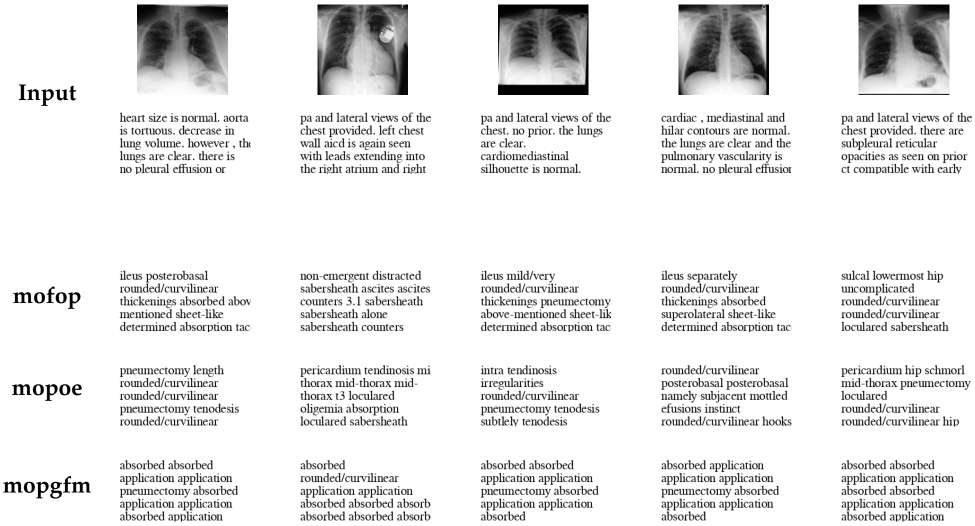

Input
heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or
pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right
pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal.
cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusior
pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early
mofop
mopoe
mopgfm

Figure B.13: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are PA, text and the generated modality is PA.



| | | | | |
|---|---|---|---|---|
| heart size is normal. aorta<br>is tortuous. decrease in<br>lung volume. however , th<br>lungs are clear. there is<br>no pleural effusion or | pa and lateral views of the<br>chest provided. left chest<br>wall aicd is again seen<br>with leads extending into<br>the right atrium and right | pa and lateral views of the<br>chest. no prior. the lungs<br>are clear.<br>cardiomediastinal<br>silhouette is normal. | cardiac , mediastinal and<br>hilar contours are normal.<br>the lungs are clear and the<br>pulmonary vascularity is<br>normal. no pleural effusio | pa and lateral views of the<br>chest provided. there are<br>subpleural reticular<br>opacities as seen on prior<br>ct compatible with early |

Figure B.14: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are PA, text and the generated modality is Lateral.



| | | | | | |
|---|---|---|---|---|---|
| Input | heart size is normal. aorta<br>is tortuous. decrease in<br>lung volume. however , th<br>lungs are clear. there is<br>no pleural effusion or | pa and lateral views of the<br>chest provided. left chest<br>wall aicd is again seen<br>with leads extending into<br>the right atrium and right | pa and lateral views of the<br>chest. no prior. the lungs<br>are clear.<br>cardiomediastinal<br>silhouette is normal. | cardiac , mediastinal and<br>hilar contours are normal.<br>the lungs are clear and the<br>pulmonary vascularity is<br>normal. no pleural effusio | pa and lateral views of the<br>chest provided. there are<br>subpleural reticular<br>opacities as seen on prior<br>ct compatible with early |
| mofop | ileus posterobasal<br>rounded/curvilinear<br>thickenings absorbed abov<br>mentioned sheet-like<br>determined absorption tac | non-emergent distracted<br>sabersheath ascites ascites<br>counters 3.1 sabersheath<br>sabersheath alone<br>sabersheath counters | ileus mild/very<br>rounded/curvilinear<br>thickenings pneumectomy<br>above-mentioned sheet-lik<br>determined absorption tac | ileus separately<br>rounded/curvilinear<br>thickenings absorbed<br>superolateral sheet-like<br>determined absorption tac | sulcal lowermost hip<br>uncomplicated<br>rounded/curvilinear<br>rounded/curvilinear<br>loculared sabersheath |
| mopoe | pneumectomy length<br>rounded/curvilinear<br>rounded/curvilinear<br>pneumectomy tenodesis<br>rounded/curvilinear | pericardium tendinosis mi<br>thorax mid-thorax mid-<br>thorax t3 loculared<br>oligemia absorption<br>loculared sabersheath | intra tendinosis<br>irregularities<br>rounded/curvilinear<br>pneumectomy tenodesis<br>subtlely tenodesis | rounded/curvilinear<br>posterobasal posterobasal<br>namely subjacent mottled<br>efusions instinct<br>rounded/curvilinear hooks | pericardium hip schmorl<br>mid-thorax pneumectomy<br>loculared<br>rounded/curvilinear<br>rounded/curvilinear hip |
| mopgfm | absorbed absorbed<br>application application<br>pneumectomy absorbed<br>application application<br>absorbed application | absorbed<br>rounded/curvilinear<br>application application<br>absorbed absorbed absorb<br>absorbed absorbed absorb | absorbed absorbed<br>application application<br>pneumectomy absorbed<br>application application<br>absorbed | absorbed application<br>application application<br>pneumectomy absorbed<br>application application<br>absorbed | absorbed absorbed<br>application application<br>absorbed absorbed<br>application application<br>absorbed application |

Figure B.15: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are PA, text and the generated modality is text.

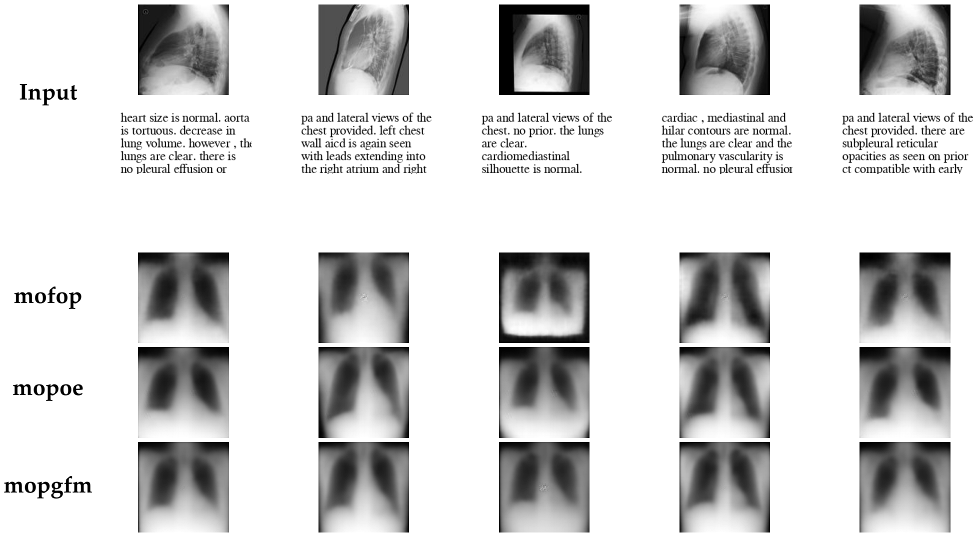

Input
heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or
pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right
pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal.
cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio
pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early
mofop
mopoe
mopgfm

Figure B.16: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, text and the generated modality is PA.

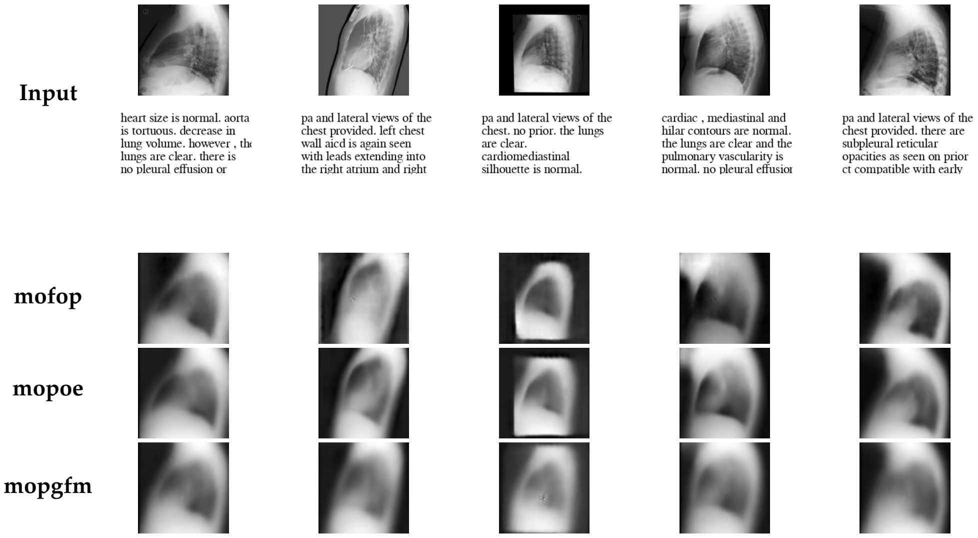
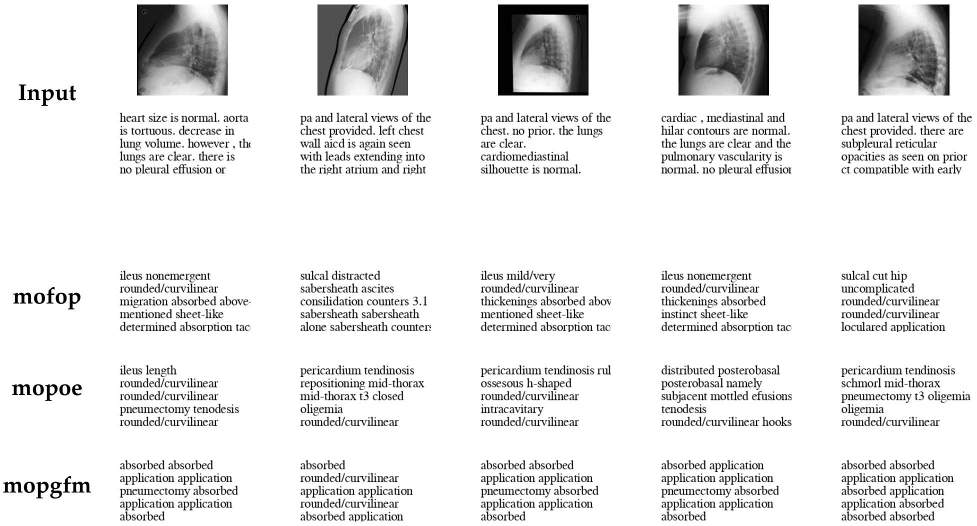
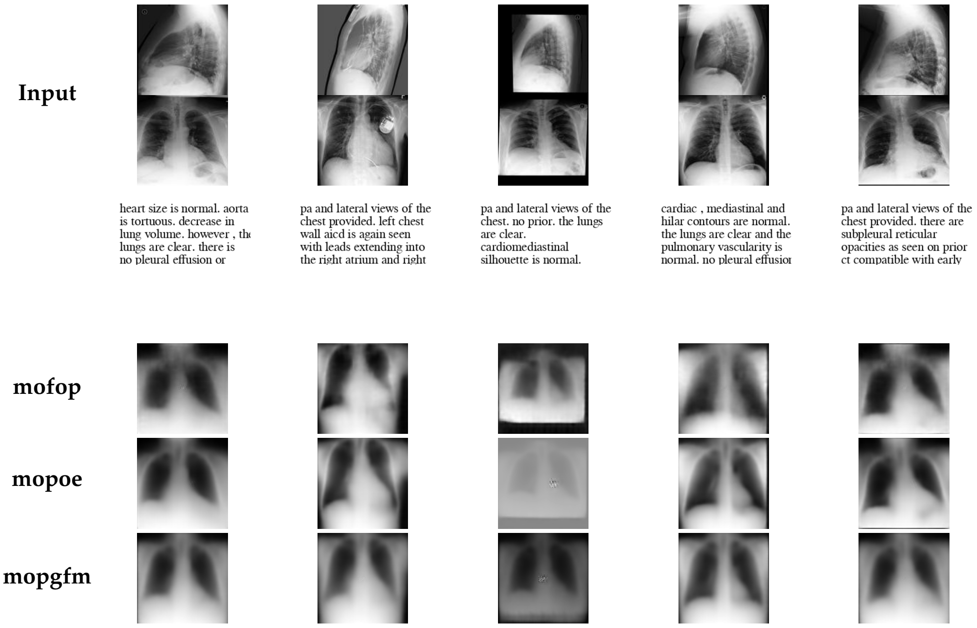

Input
heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or
pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right
pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal.
cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio
pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early
mofop
mopoe
mopgfm

Figure B.17: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, text and the generated modality is Lateral.



| | | | | | |
|---|---|---|---|---|---|
| Input | heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or | pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right | pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal. | cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio | pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early |
| mofop | ileus nonemergent rounded/curvilinear migration absorbed above-mentioned sheet-like determined absorption tac | sulcal distracted sabersheath ascites consilidation counters 3.1 sabersheath sabersheath alone sabersheath counter | ileus mild/very rounded/curvilinear thickenings absorbed abov mentioned sheet-like determined absorption tac | ileus nonemergent rounded/curvilinear thickenings absorbed instinct sheet-like determined absorption tac | sulcal cut hip uncomplicated rounded/curvilinear rounded/curvilinear loculared application |
| mopoe | ileus length rounded/curvilinear rounded/curvilinear pneumectomy tenodesis rounded/curvilinear | pericardium tendinosis repositioning mid-thorax mid-thorax t3 closed oligemia rounded/curvilinear | pericardium tendinosis rul ossesous h-shaped rounded/curvilinear intracavitary rounded/curvilinear | distributed posterobasal posterobasal namely subjacent mottled efusions tenodesis rounded/curvilinear hooks | pericardium tendinosis schmorl mid-thorax pneumectomy t3 oligemia oligemia rounded/curvilinear |
| mopgfm | absorbed absorbed application application pneumectomy absorbed application application absorbed | absorbed rounded/curvilinear application application rounded/curvilinear absorbed application | absorbed absorbed application application pneumectomy absorbed application application absorbed | absorbed application application application pneumectomy absorbed application application absorbed | absorbed absorbed application application absorbed application application absorbed absorbed absorbed |

Figure B.18: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, text and the generated modality is text.



| | | | | | |
|---|---|---|---|---|---|
| Input | heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or | pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right | pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal. | cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio | pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early |

Figure B.19: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA, text and the generated modality is PA.

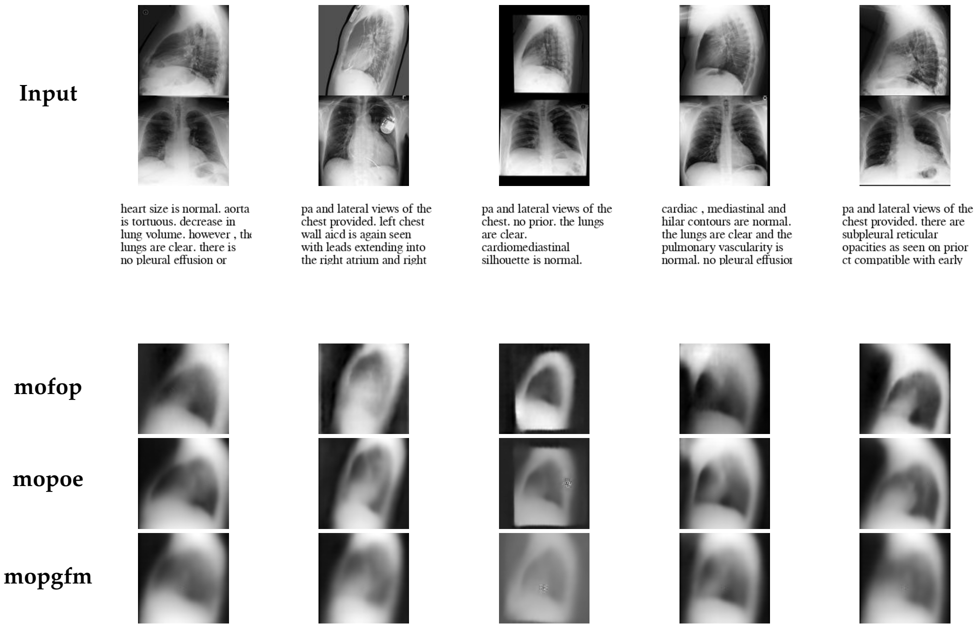

Input
heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or
pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right
pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal.
cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio
pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early
mofop
mopoe
mopgfm

Figure B.20: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA, text and the generated modality is Lateral.

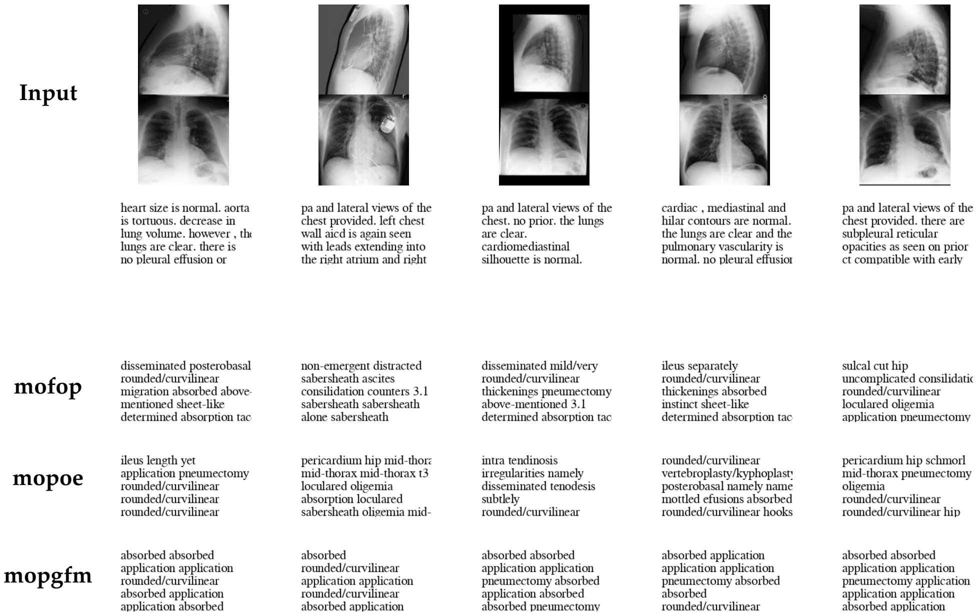

Input
heart size is normal. aorta is tortuous. decrease in lung volume. however , th lungs are clear. there is no pleural effusion or
pa and lateral views of the chest provided. left chest wall aicd is again seen with leads extending into the right atrium and right
pa and lateral views of the chest. no prior. the lungs are clear. cardiomediastinal silhouette is normal.
cardiac , mediastinal and hilar contours are normal. the lungs are clear and the pulmonary vascularity is normal. no pleural effusio
pa and lateral views of the chest provided. there are subpleural reticular opacities as seen on prior ct compatible with early
mofop
disseminated posterobasal rounded/curvilinear migration absorbed above-mentioned sheet-like determined absorption tac
non-emergent distracted sabersheath ascites consilidation counters 3.1 sabersheath sabersheath alone sabersheath
disseminated mild/very rounded/curvilinear thickenings pneumectomy above-mentioned 3.1 determined absorption tac
ileus separately rounded/curvilinear thickenings absorbed instinct sheet-like determined absorption tac
sulcal cut hip uncomplicated consilidatio rounded/curvilinear loculared oligemia application pneumectomy
mopoe
ileus length yet application pneumectomy rounded/curvilinear rounded/curvilinear rounded/curvilinear
pericardium hip mid-thora mid-thorax mid-thorax t3 loculared oligemia absorption loculared sabersheath oligemia mid-
intra tendinosis irregularities namely disseminated tenodesis subtlely rounded/curvilinear
rounded/curvilinear vertebroplasty/kyphoplast posterobasal namely name mottled efusions absorbed rounded/curvilinear hooks
pericardium hip schmorl mid-thorax pneumectomy oligemia rounded/curvilinear rounded/curvilinear hip
mopgfm
absorbed absorbed application application rounded/curvilinear absorbed application application absorbed
absorbed rounded/curvilinear application application rounded/curvilinear absorbed application
absorbed absorbed application application pneumectomy absorbed application absorbed absorbed pneumectomy
absorbed application application application pneumectomy absorbed absorbed rounded/curvilinear
absorbed absorbed application application pneumectomy application application application absorbed application

Figure B.21: Generated examples, conditioned on samples from the MIMIC-CXR Test set. The input modalities are Lateral, PA, text and the generated modality is text.

Appendix C

# Qualitative comparison across different number of importance samples

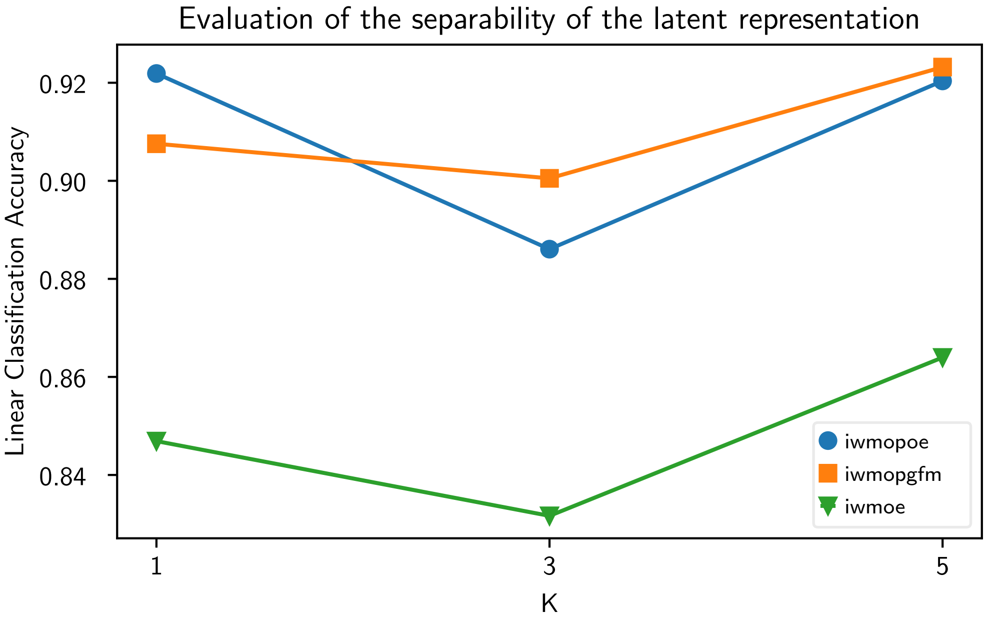


Figure C.1: **Linear classification accuracy for different importance samples over the PolyMNIST test set, averaged over all subsets.** All methods were trained with 3 modalities.

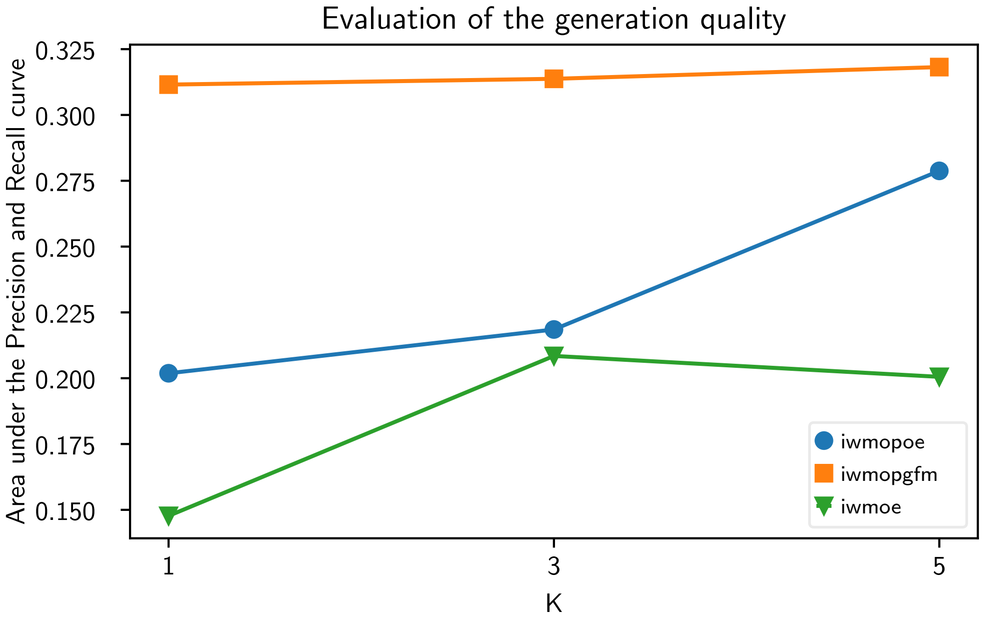


Figure C.2: **Area under the Precision and Recall curve of the PRD metric (Sajjadi et al., 2018), evaluated on the PolyMNIST test set.** All methods were trained with 3 modalities.

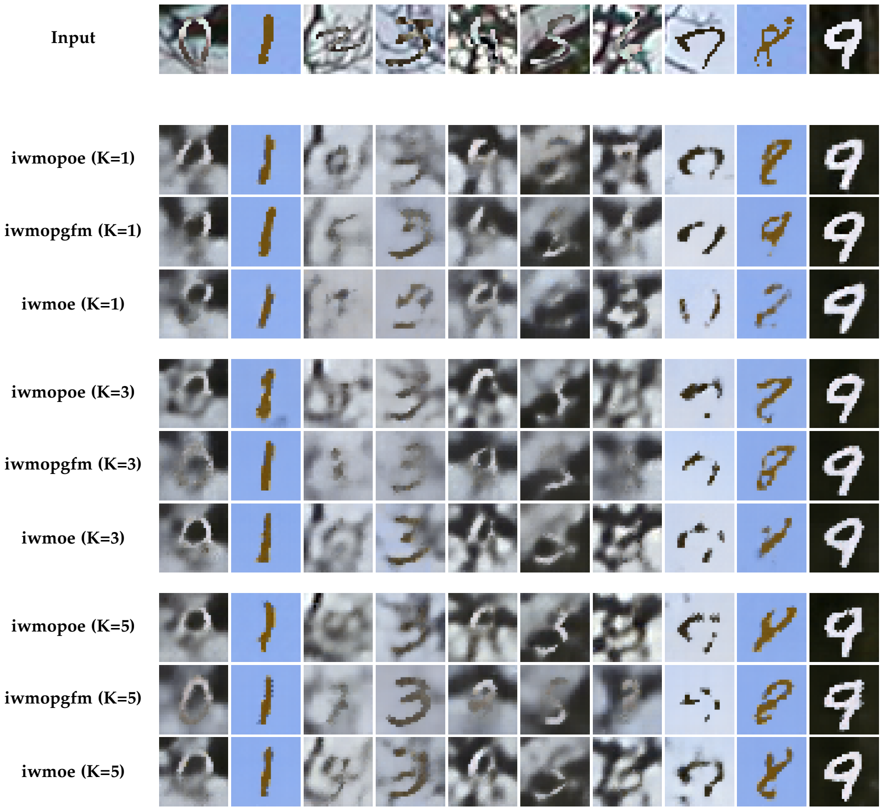


Figure C.3: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m0.

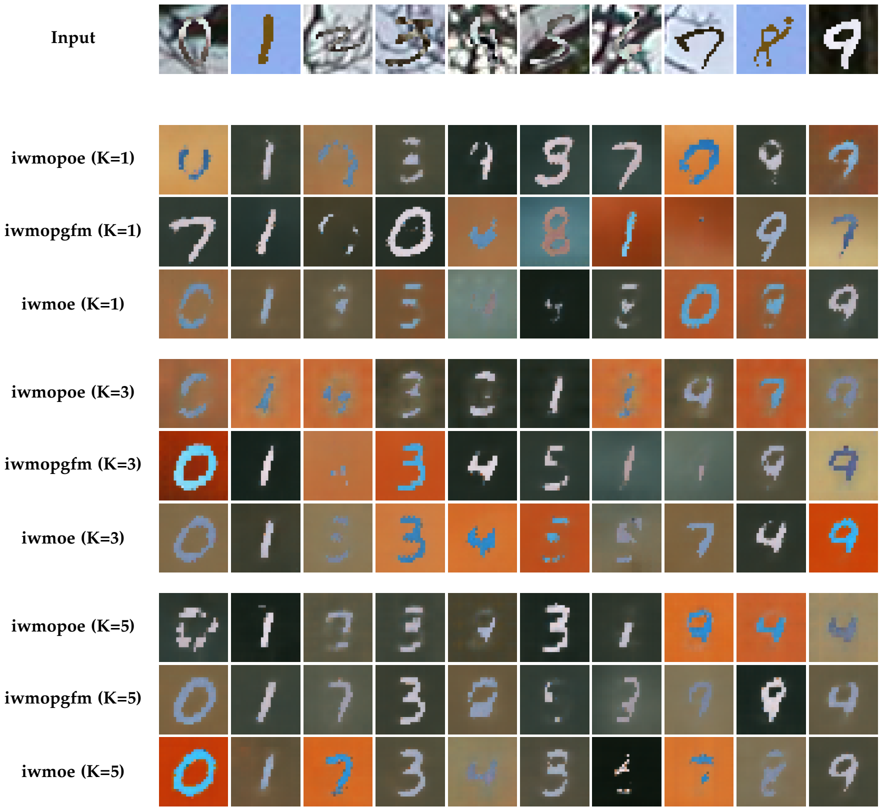


Figure C.4: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m1.

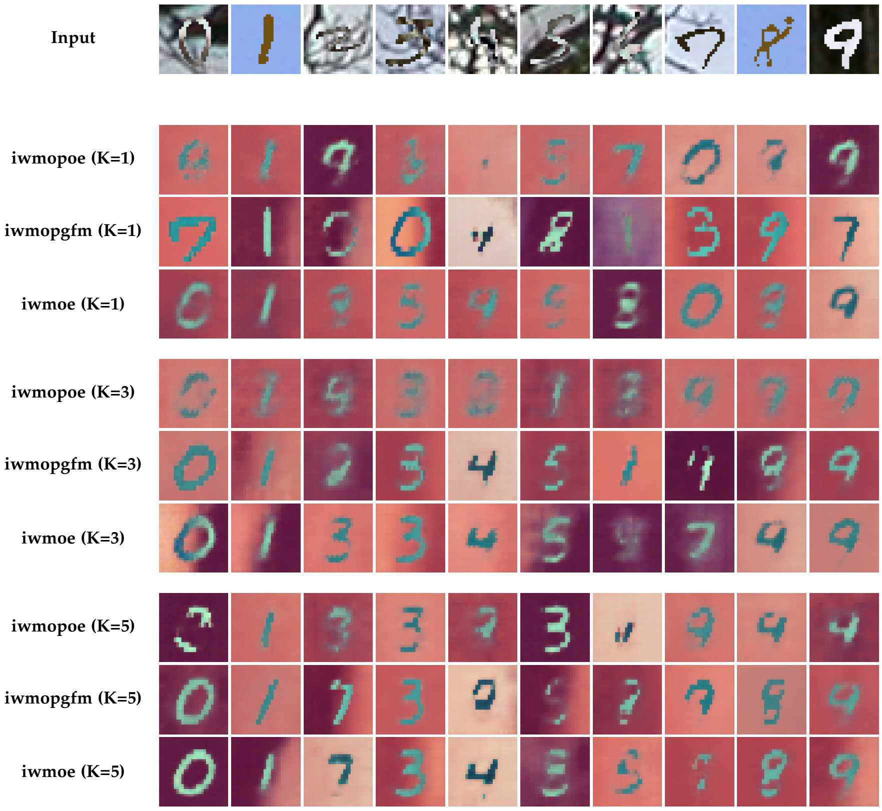


Figure C.5: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m0 and the generated modality is m2.

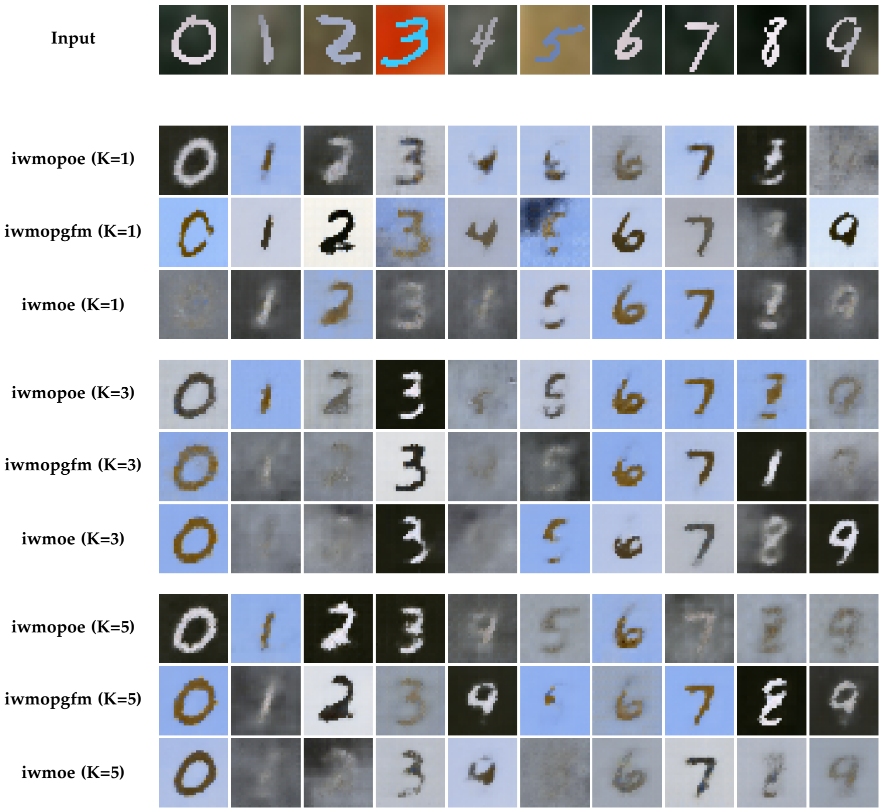


Figure C.6: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m0.

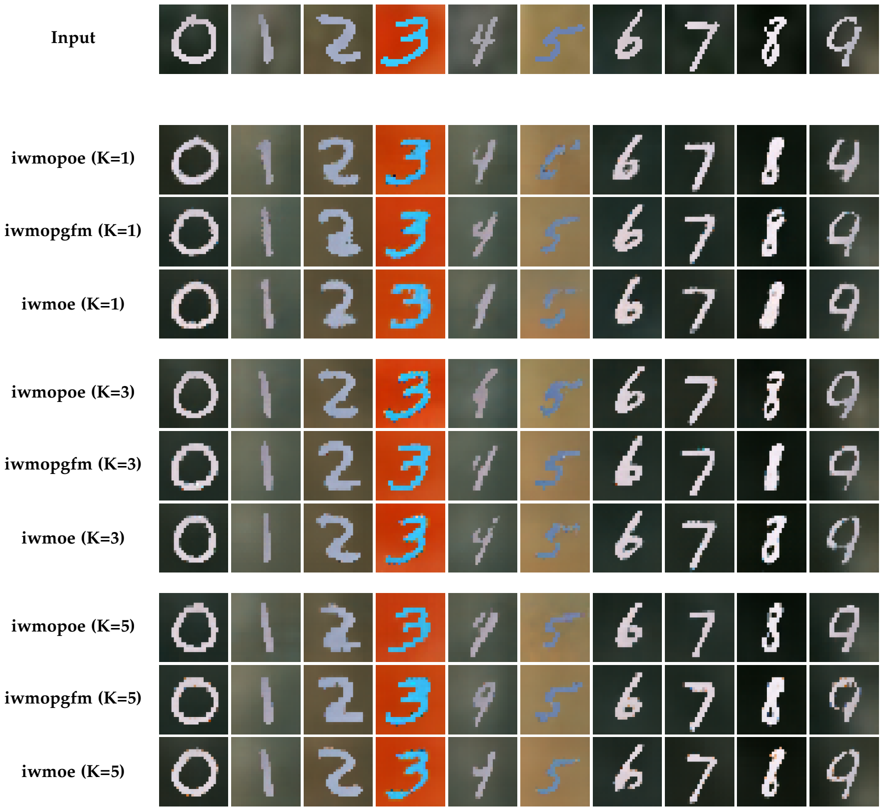


Figure C.7: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m1.

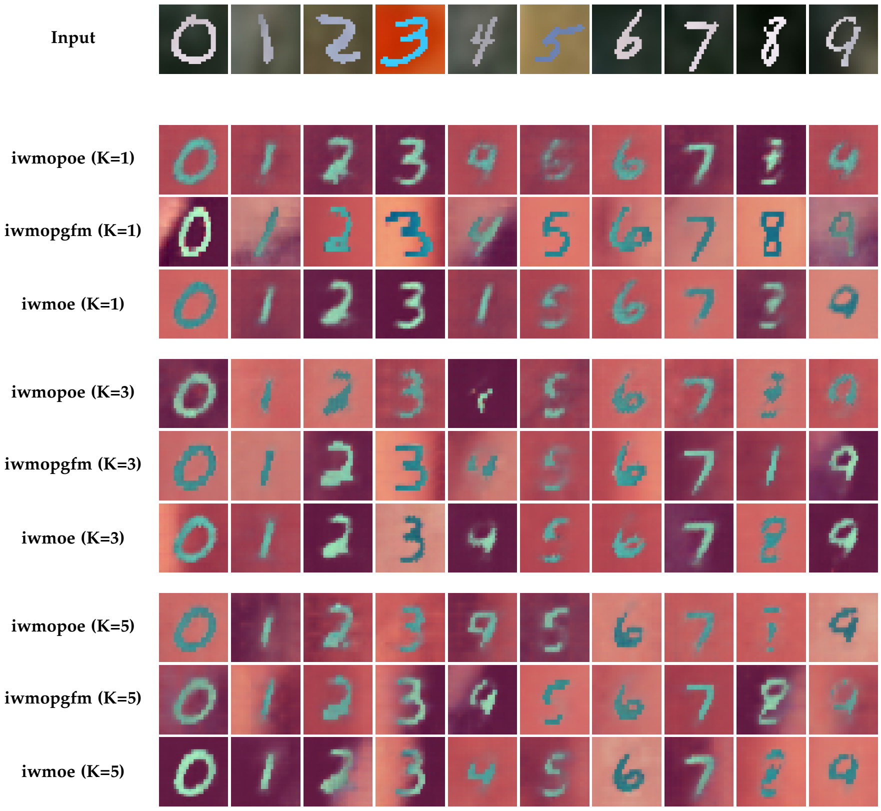


Figure C.8: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m1 and the generated modality is m2.

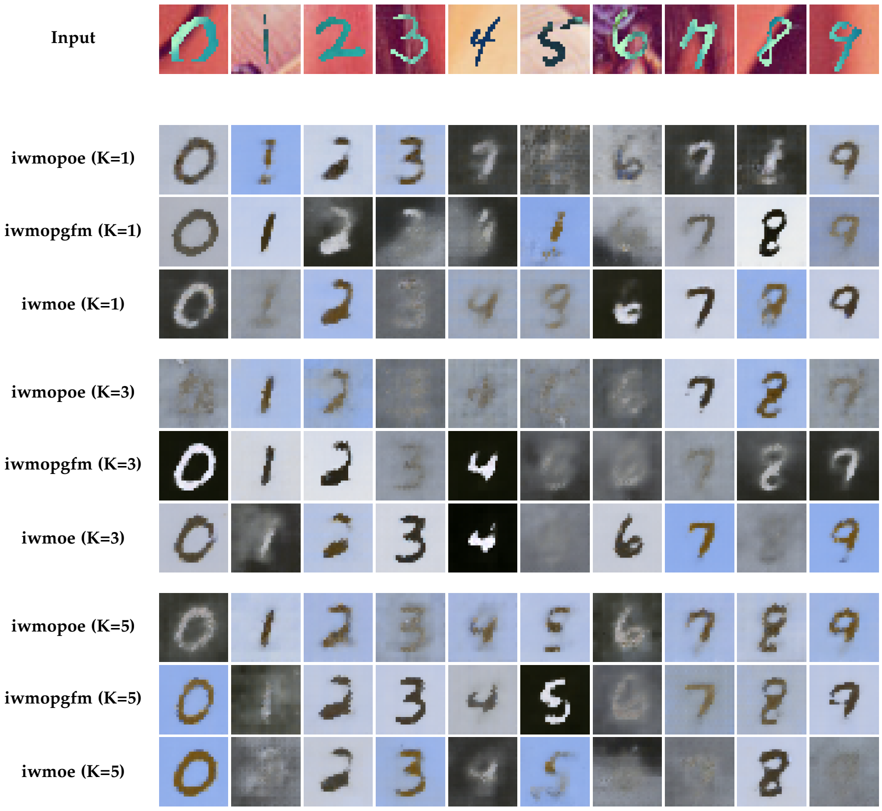


Figure C.9: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m0.

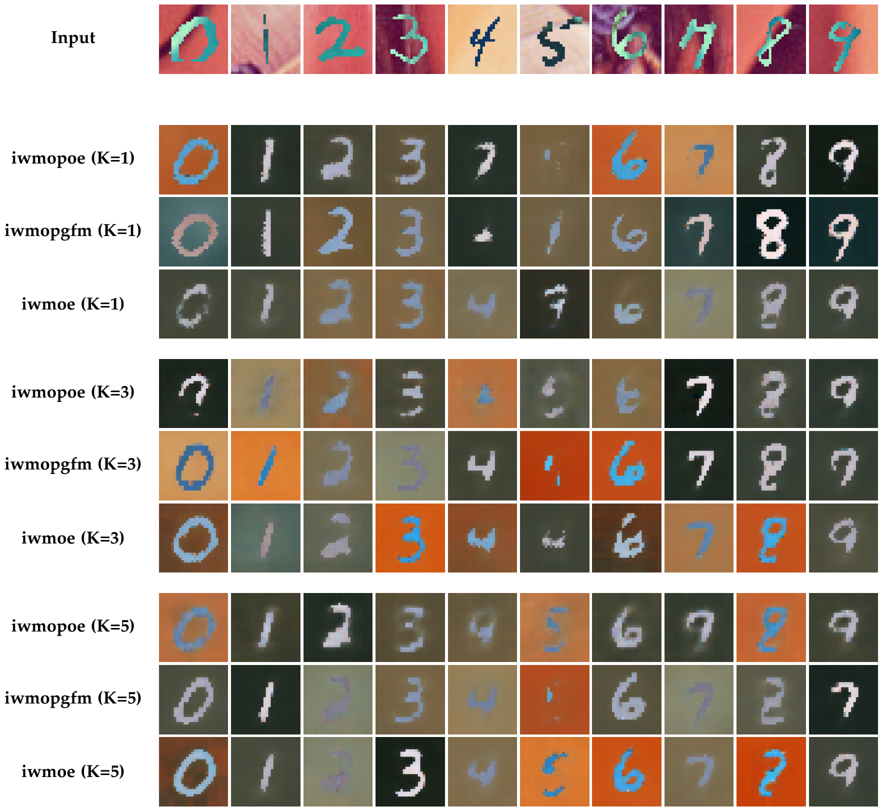


Figure C.10: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m1.

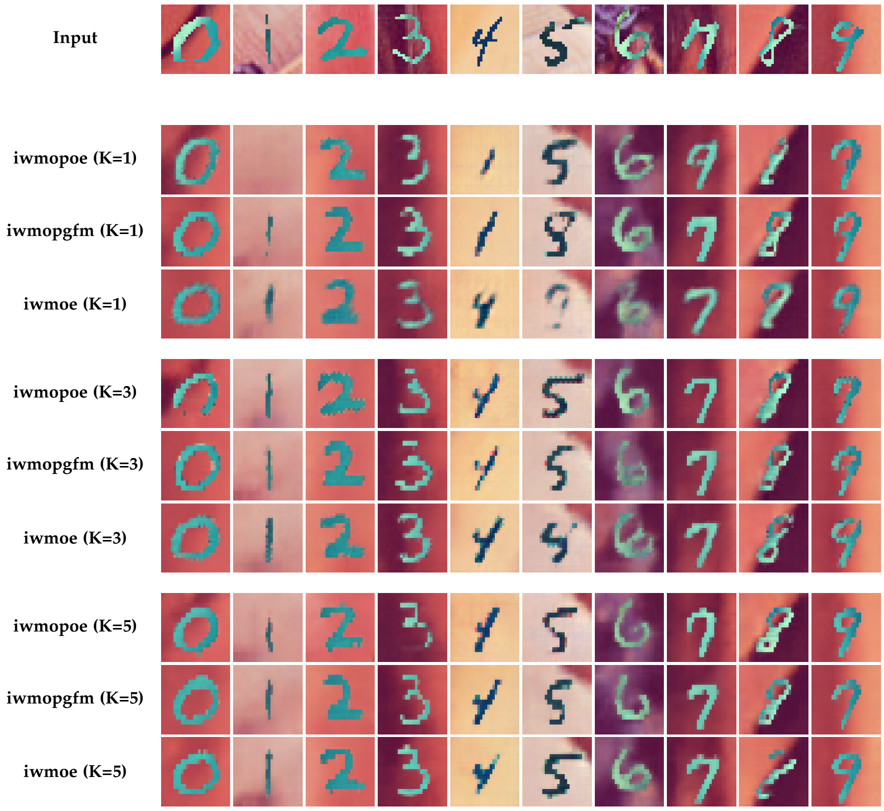


Figure C.11: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modality is m2 and the generated modality is m2.

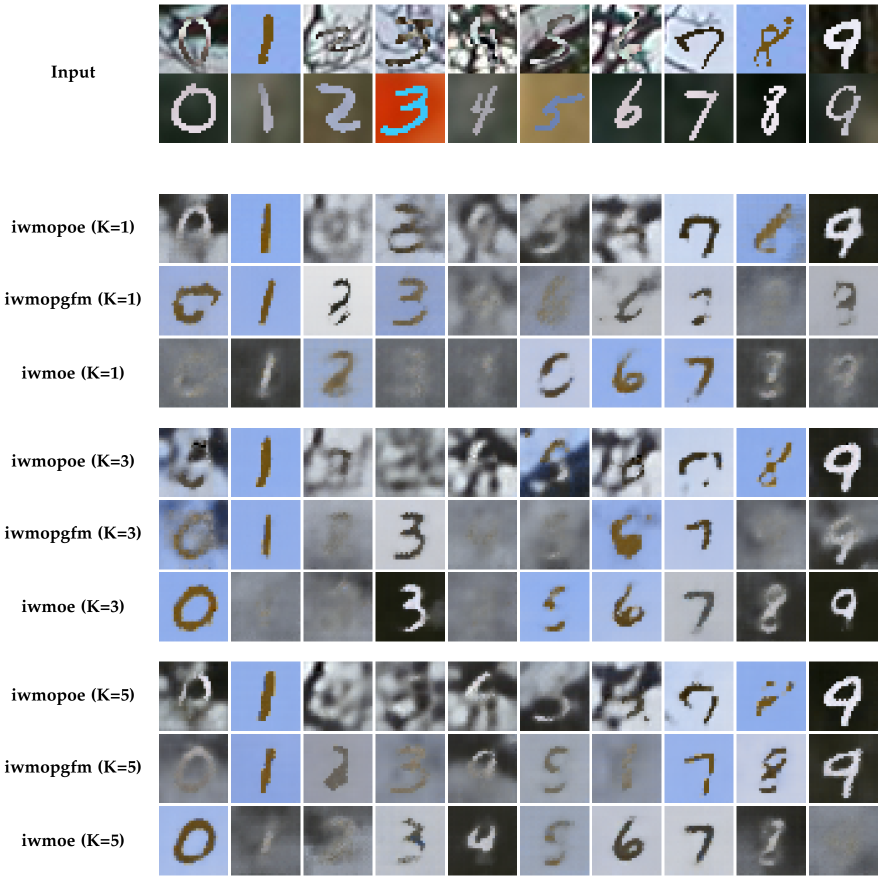


Figure C.12: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m0.

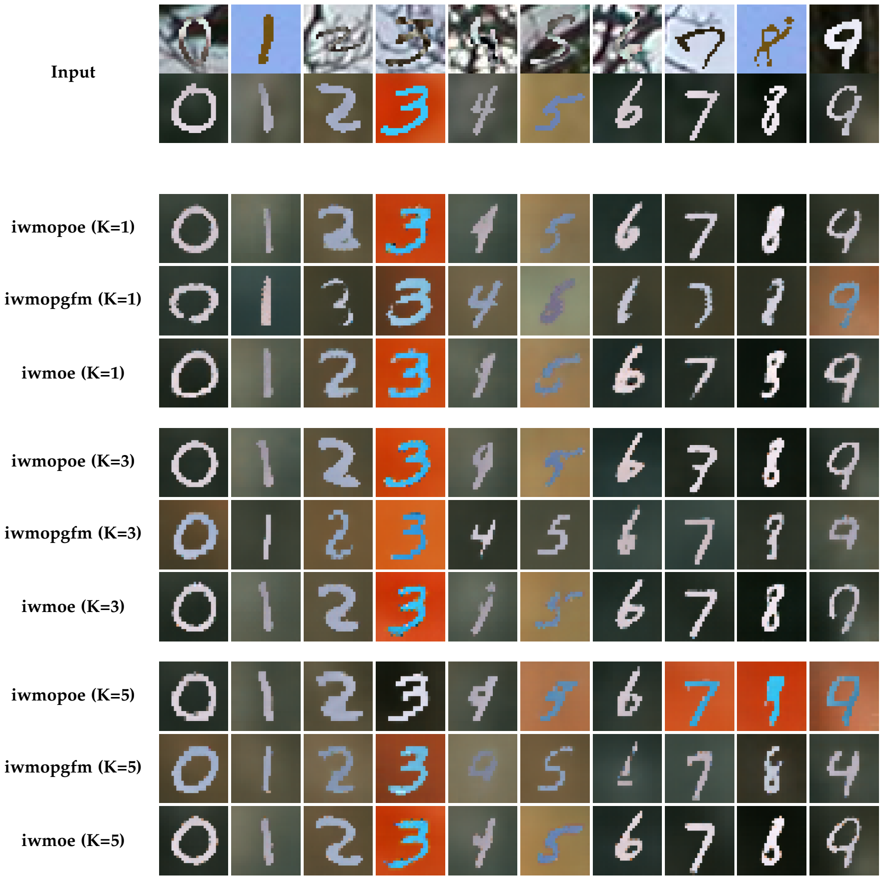


Figure C.13: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m1.

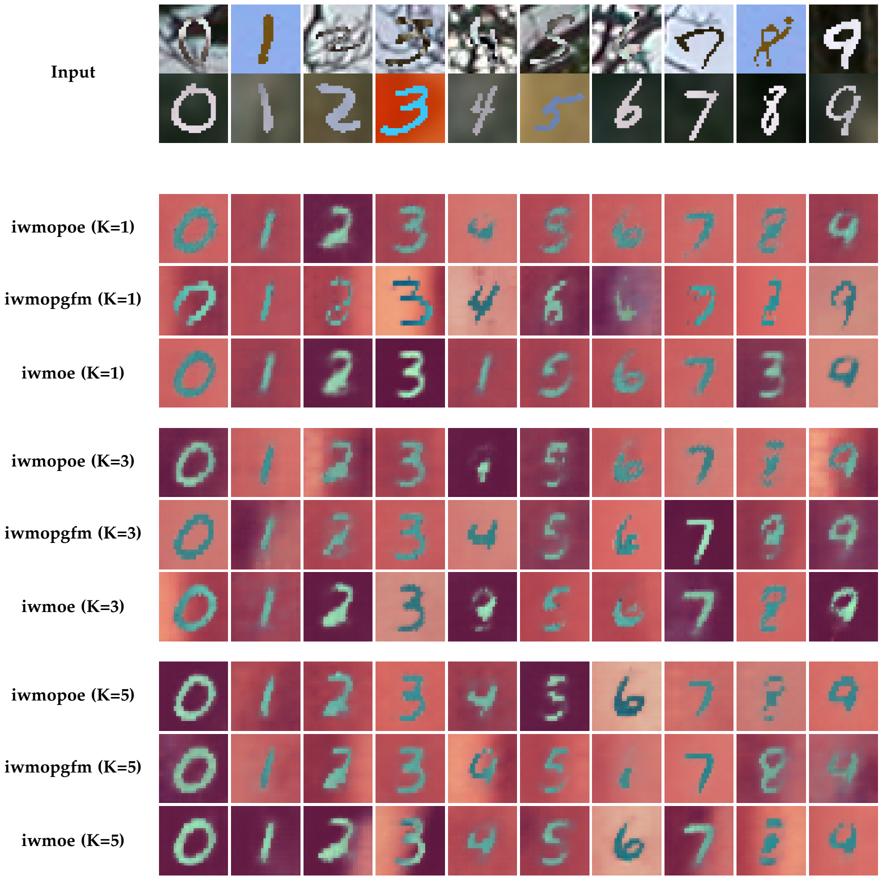


Figure C.14: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1 and the generated modality is m2.

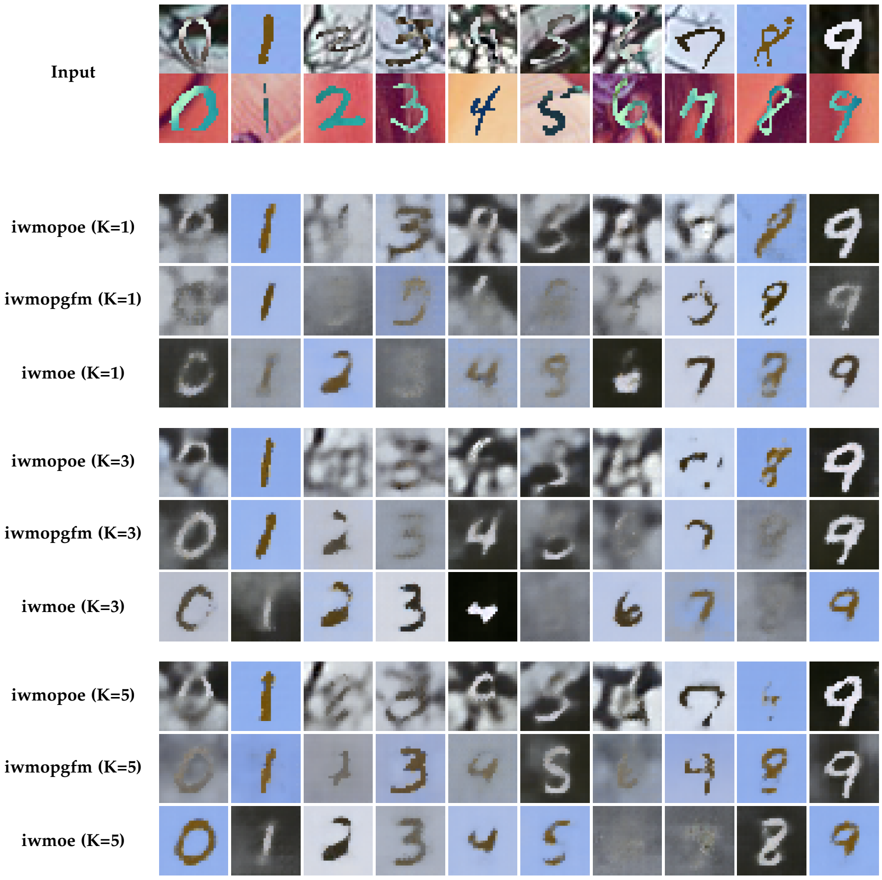


Figure C.15: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m0.

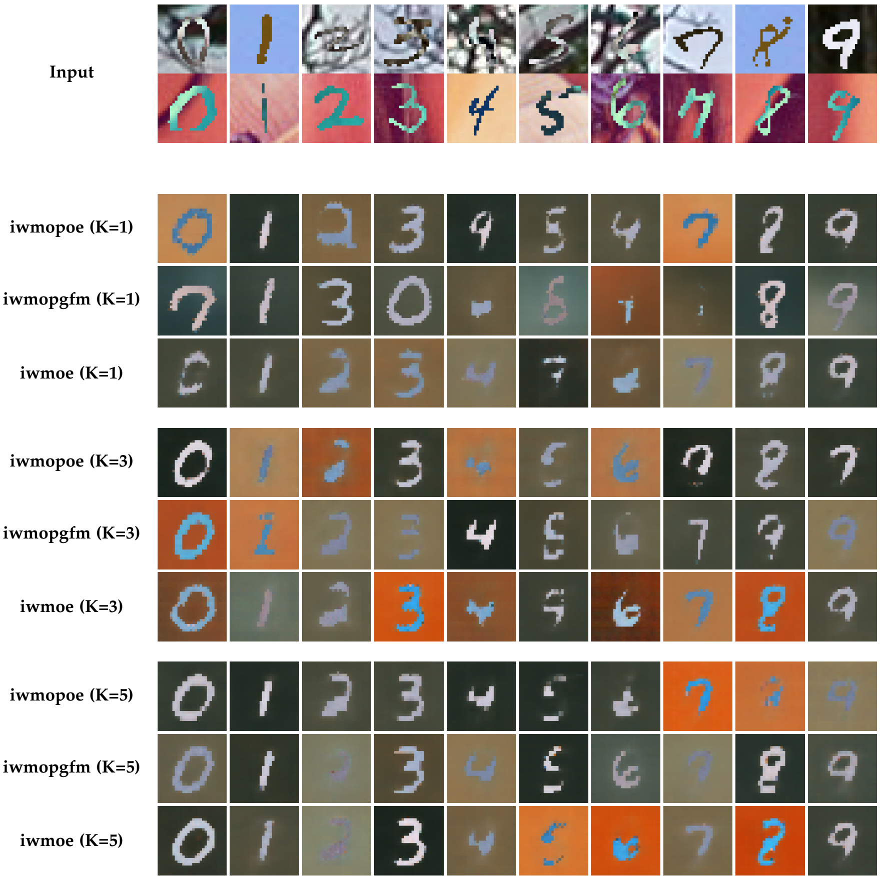


Figure C.16: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m1.

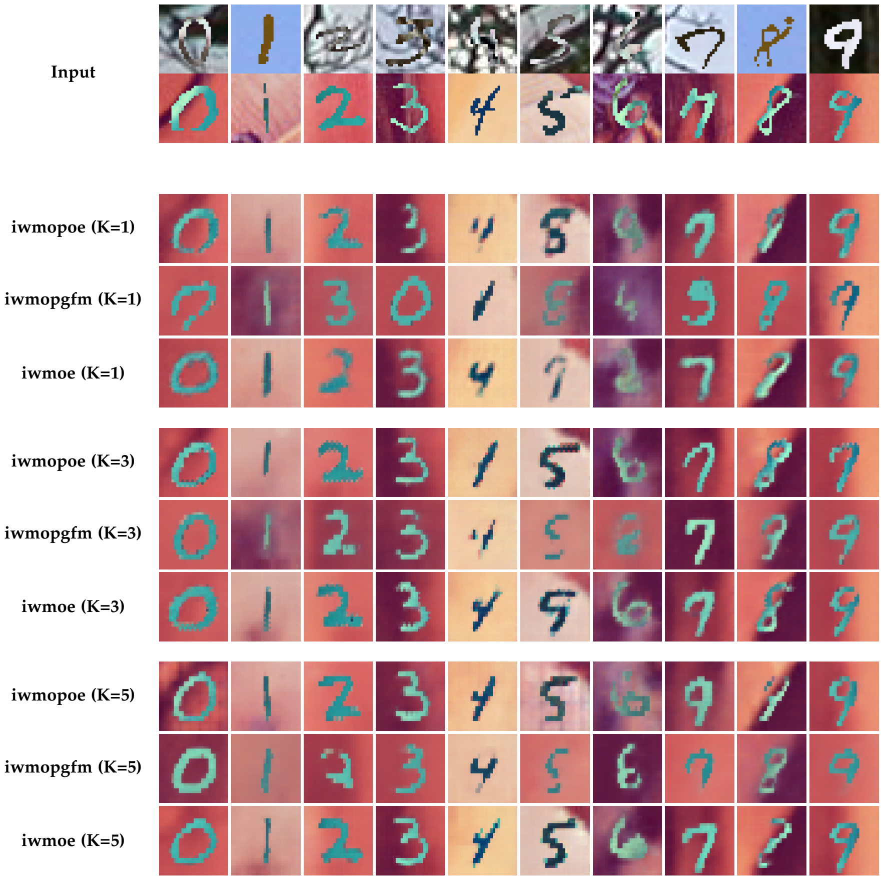


Figure C.17: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m2 and the generated modality is m2.

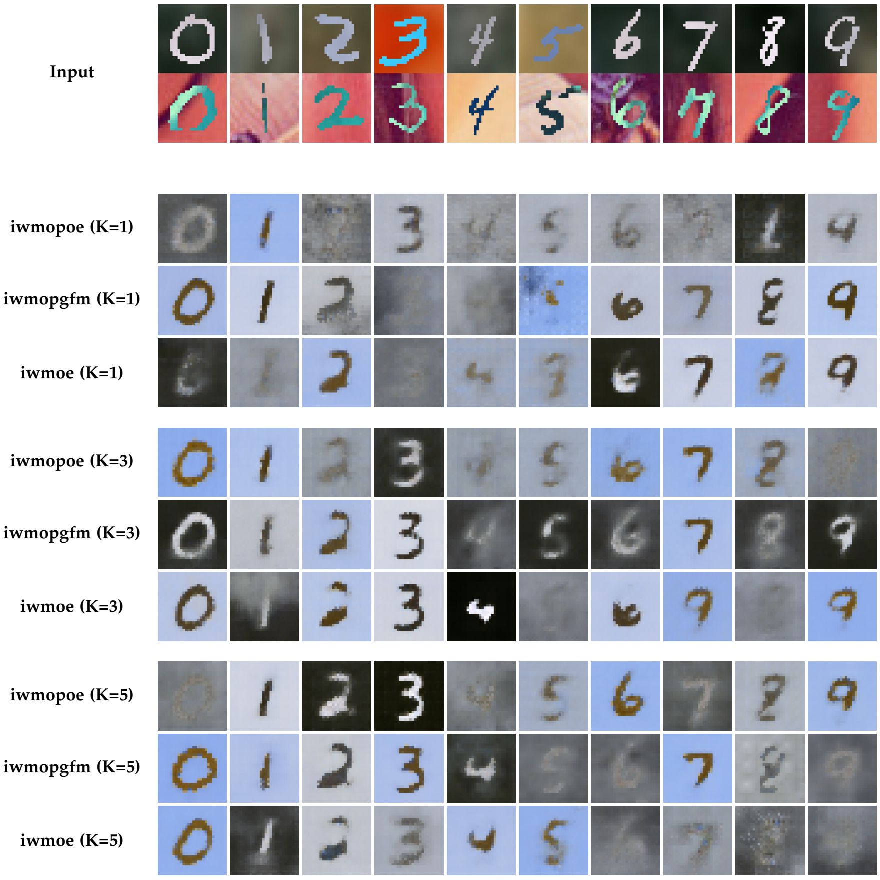


Figure C.18: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m0.

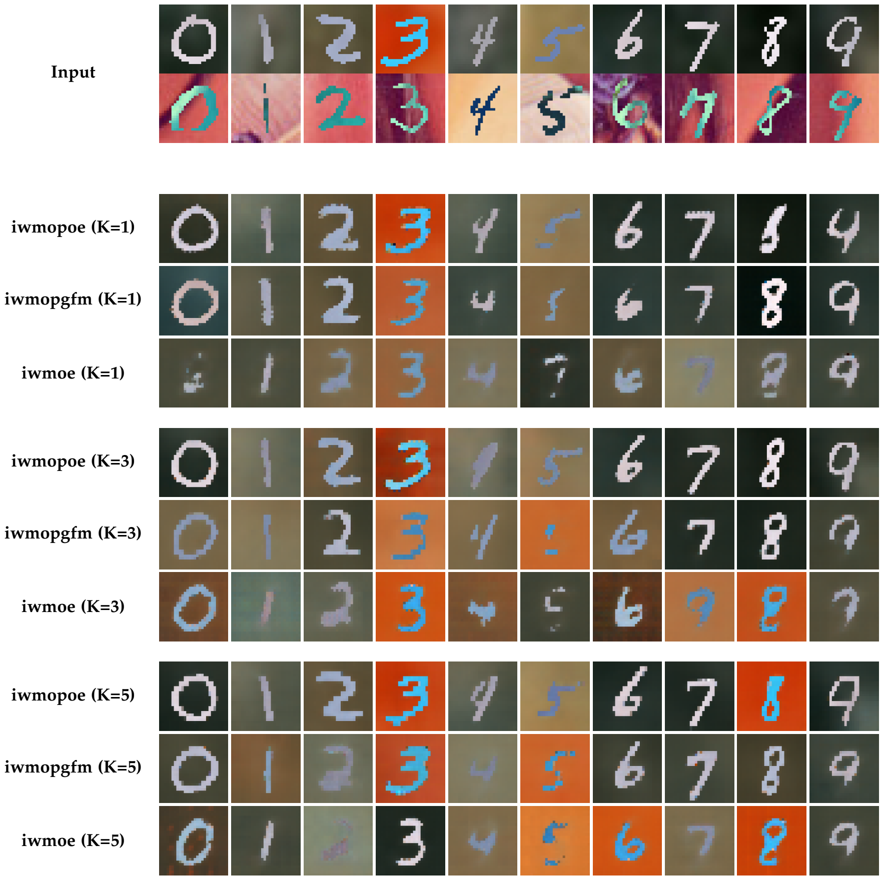


Figure C.19: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m1.

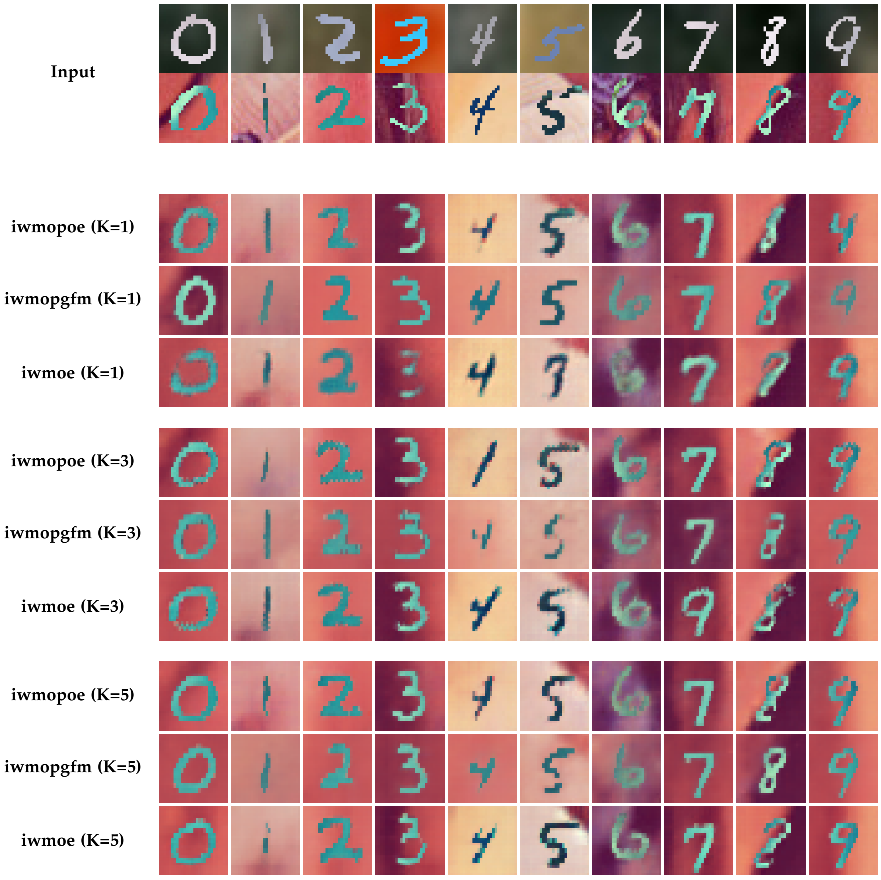


Figure C.20: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m1, m2 and the generated modality is m2.

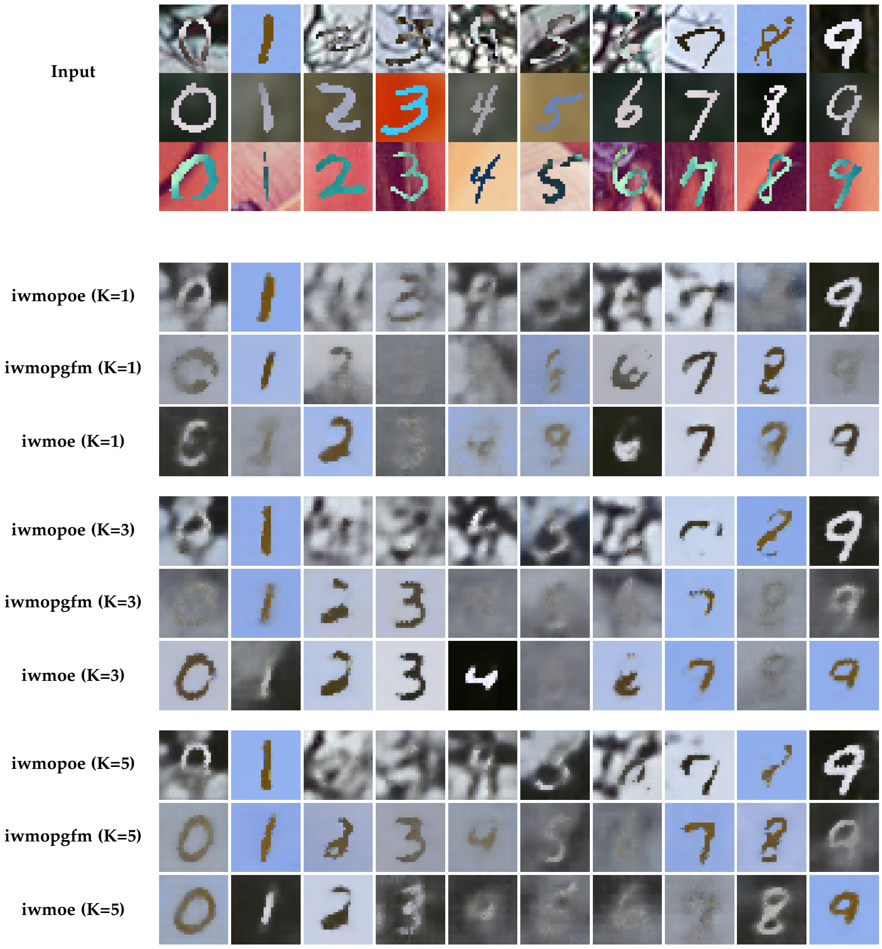


Figure C.21: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m0.

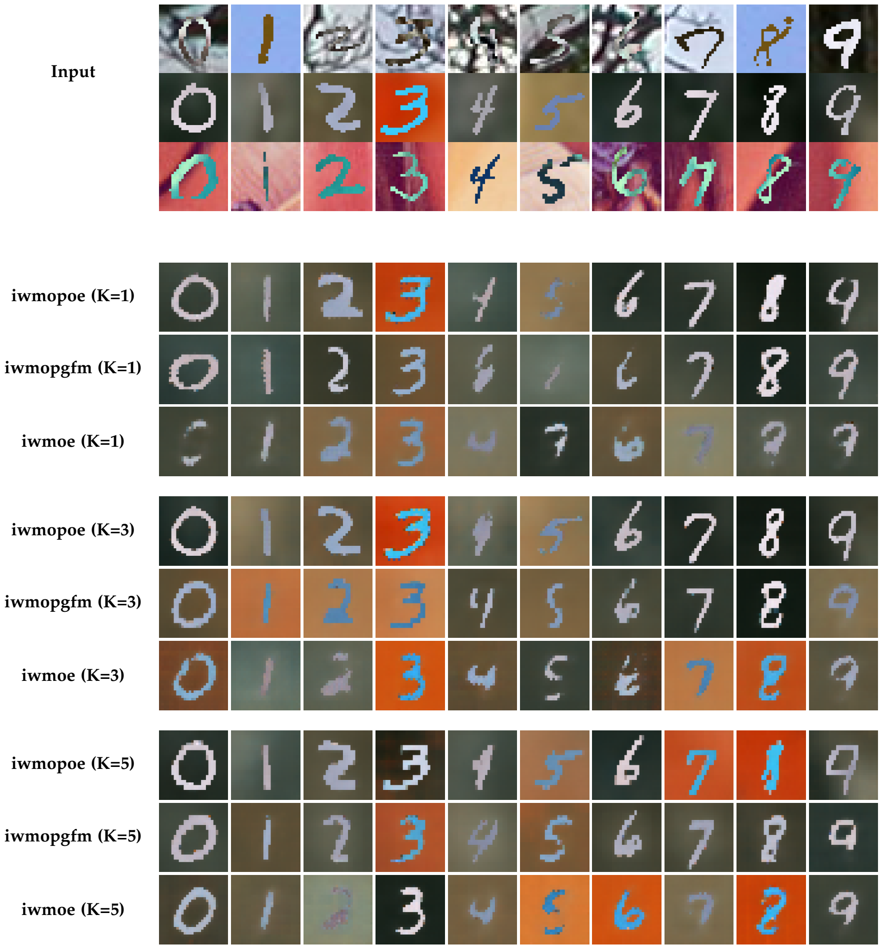


Figure C.22: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m1.

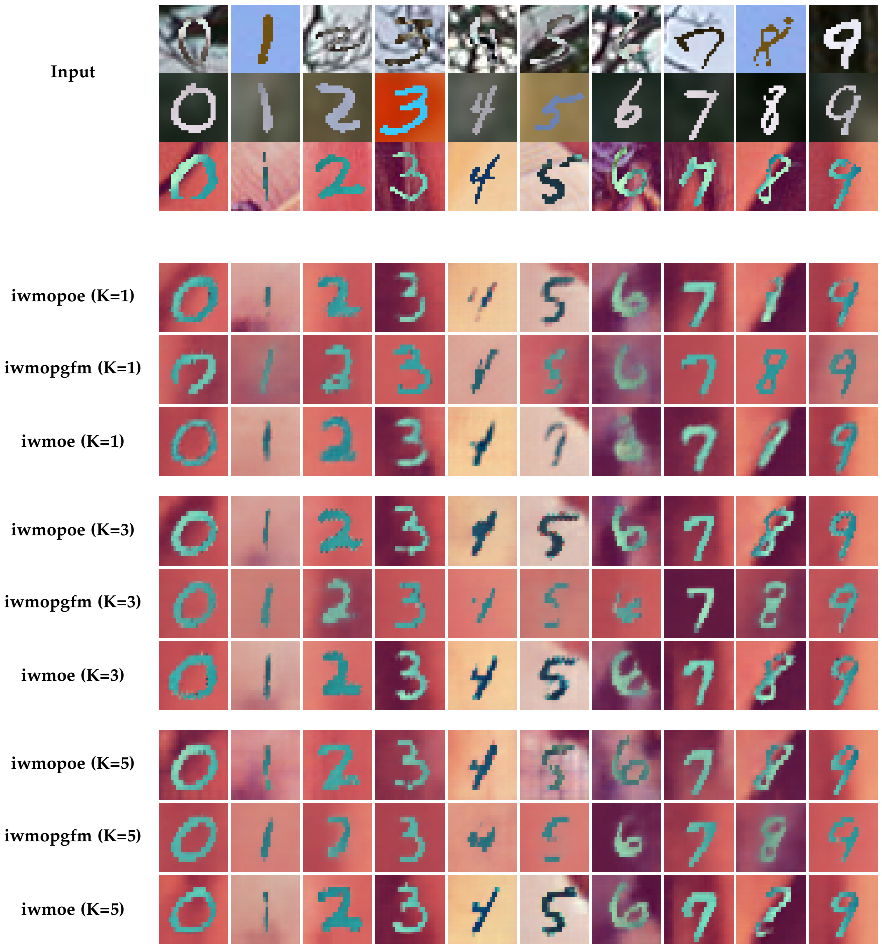


Figure C.23: Generated examples with different number of important samples (K), conditioned on samples from the PolyMNIST Test set. The input modalities are m0, m1, m2 and the generated modality is m2.