MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

European Research Council
Established by the European Commission

UNIVERSITÄT
DES
SAARLANDES

AISTATS

Hellinger Multimodal Variational Autoencoders

Presenter: Huyen Vo

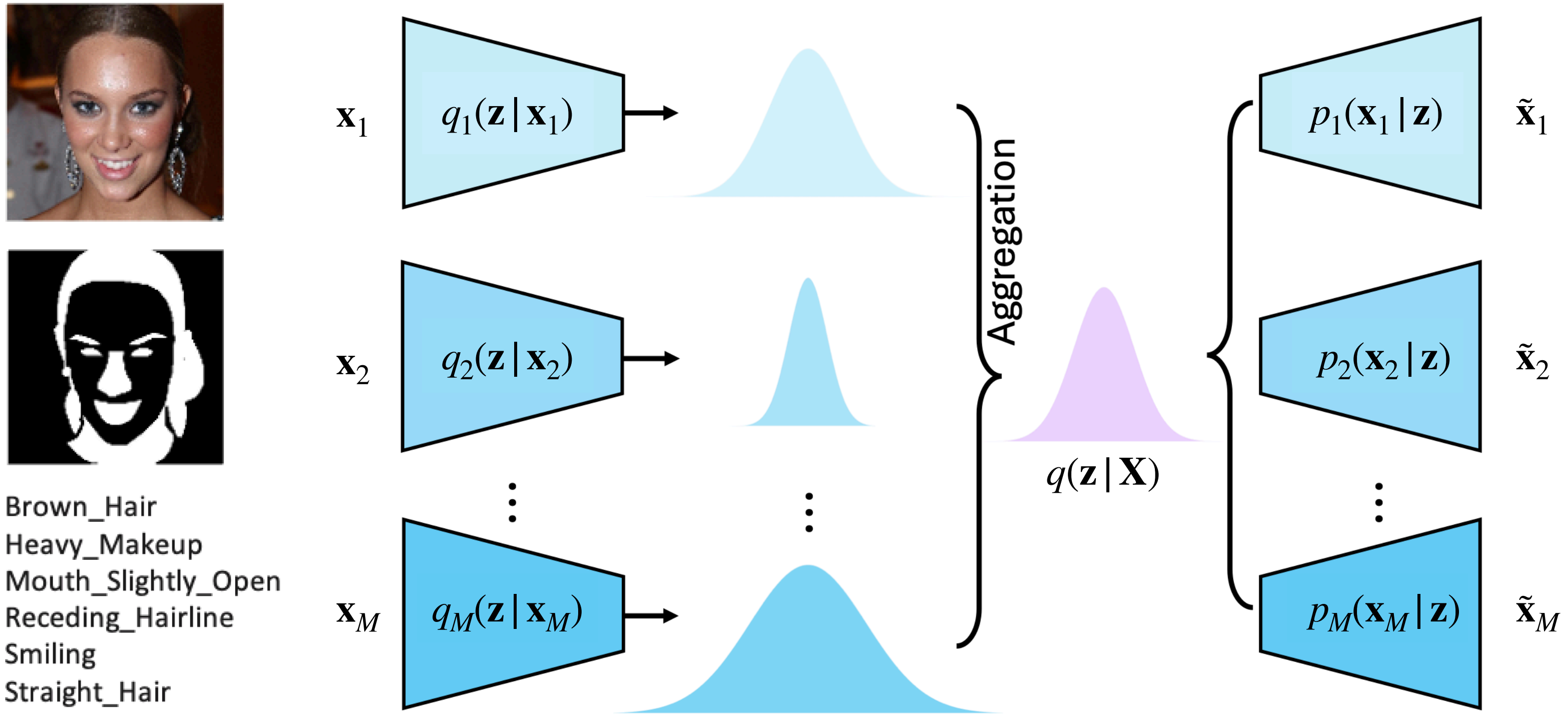
AISTATS 2026 - Spotlight

Huyen Vo

Isabel Valera

Motivation

In **multimodal generative learning**



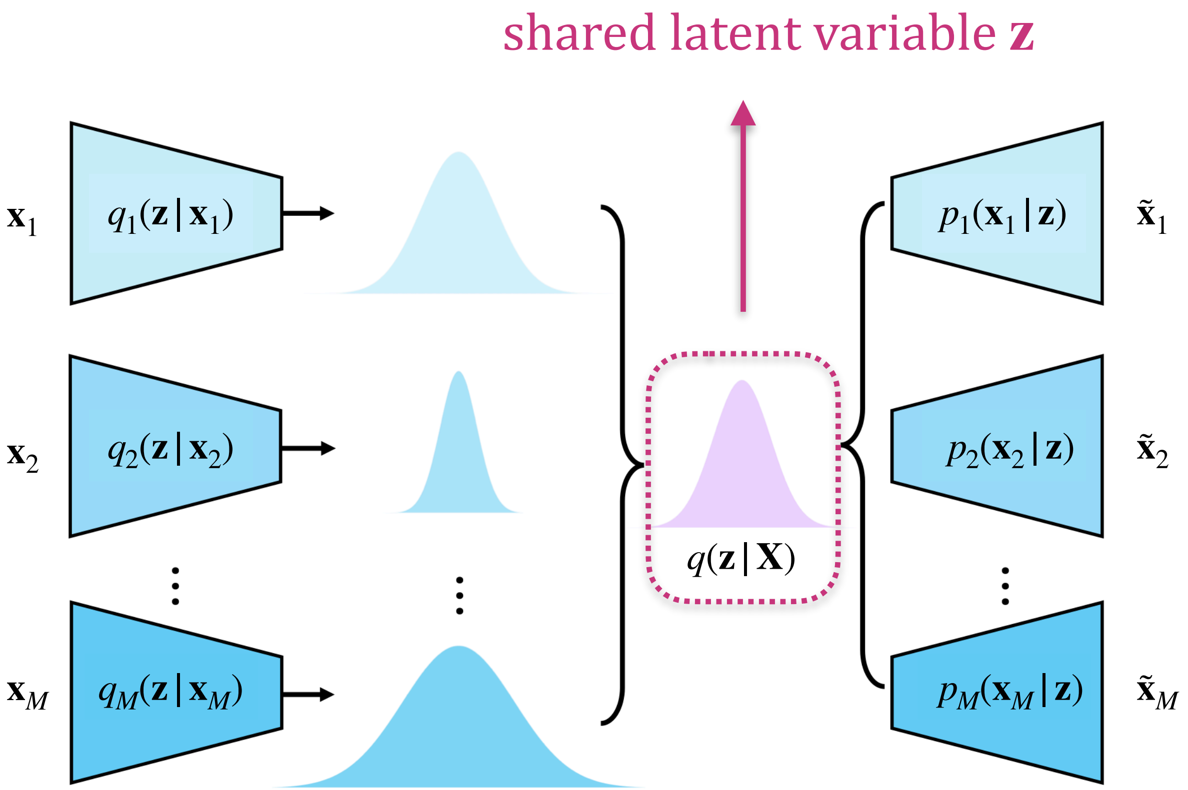
[1]

[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

We want **multimodal generative models** that support

1. learn shared representations,



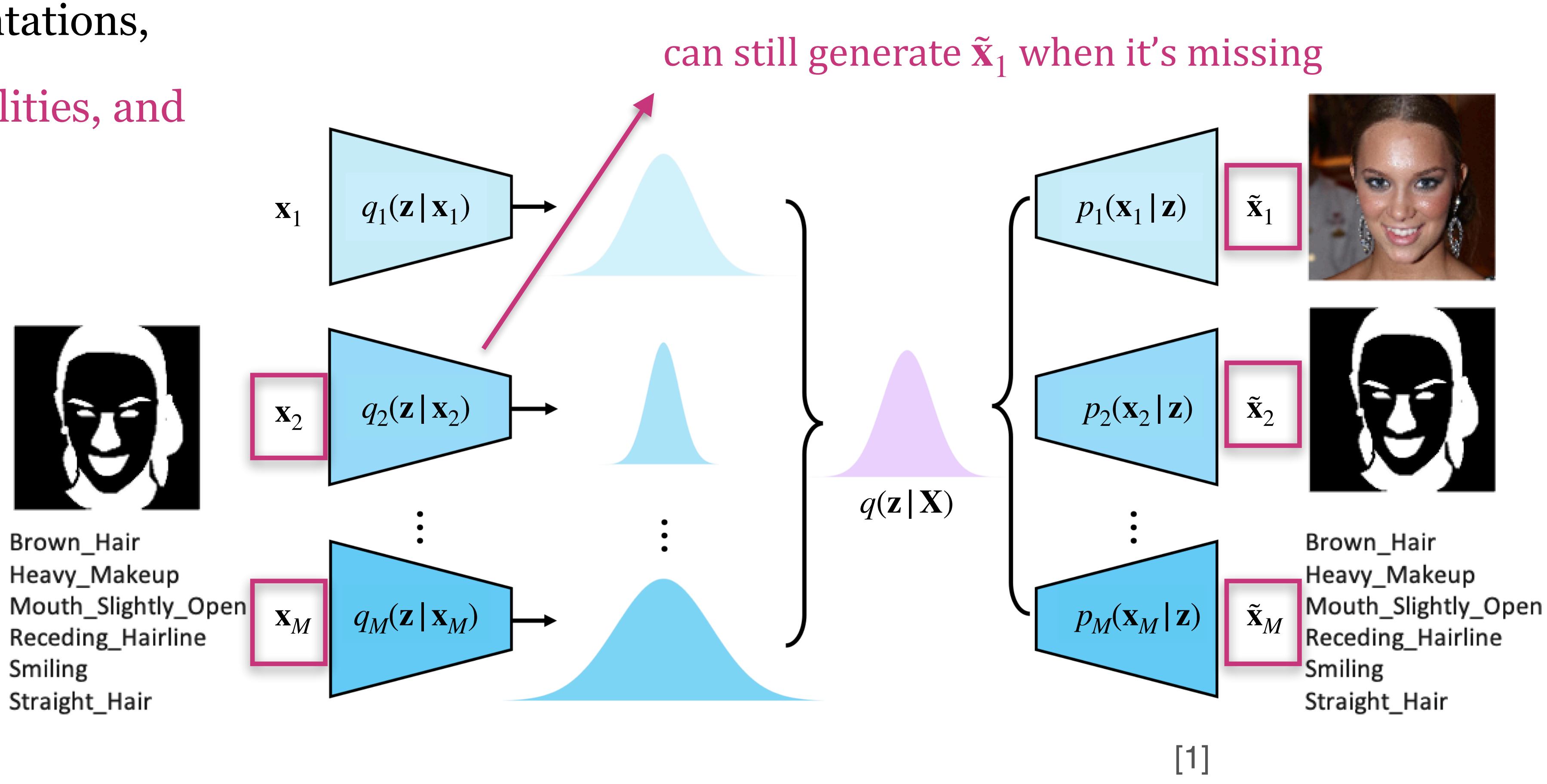
[1]

[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

We want **multimodal generative models** that support

1. learn shared representations,
2. handle missing modalities, and

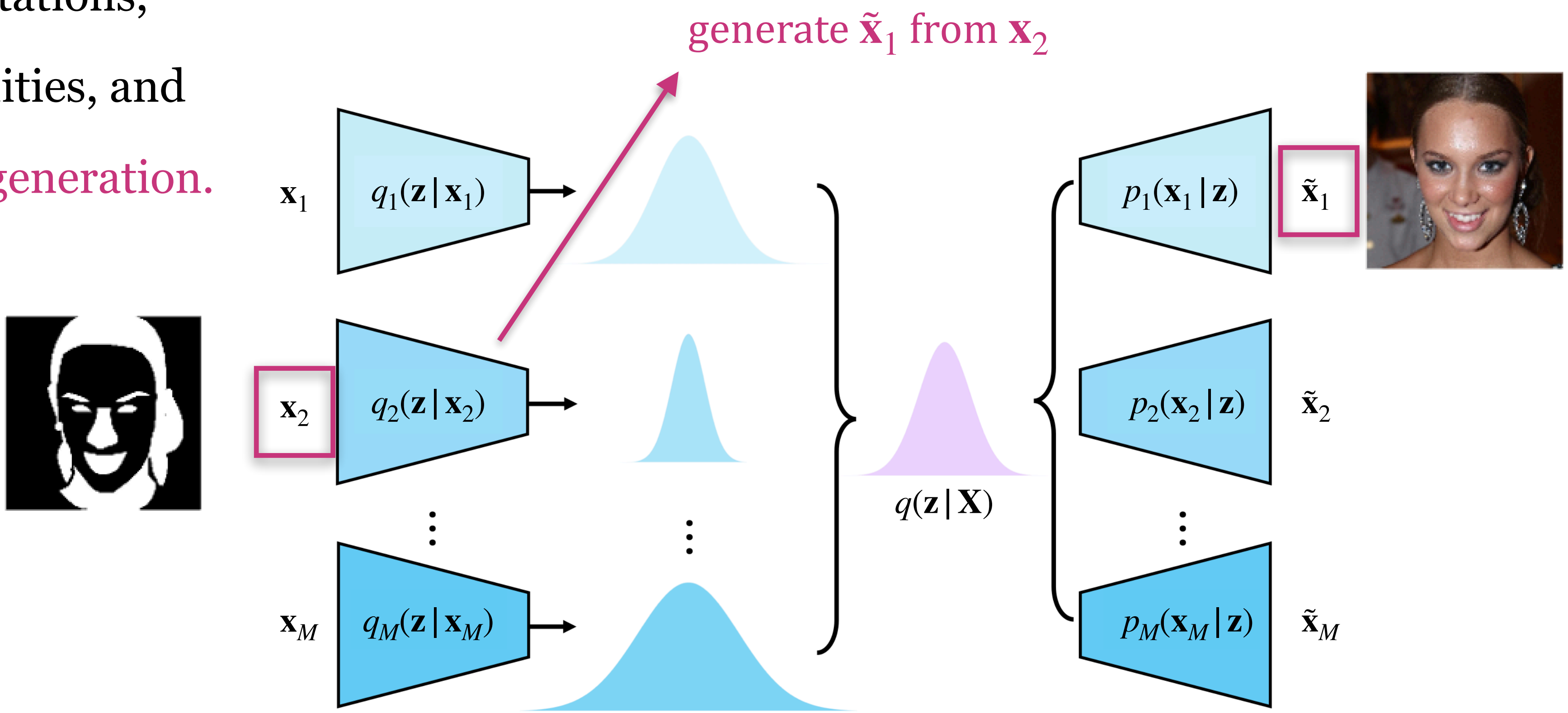


[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

We want **multimodal generative models** that support

1. learn shared representations,
2. handle missing modalities, and
3. support cross-modal generation.

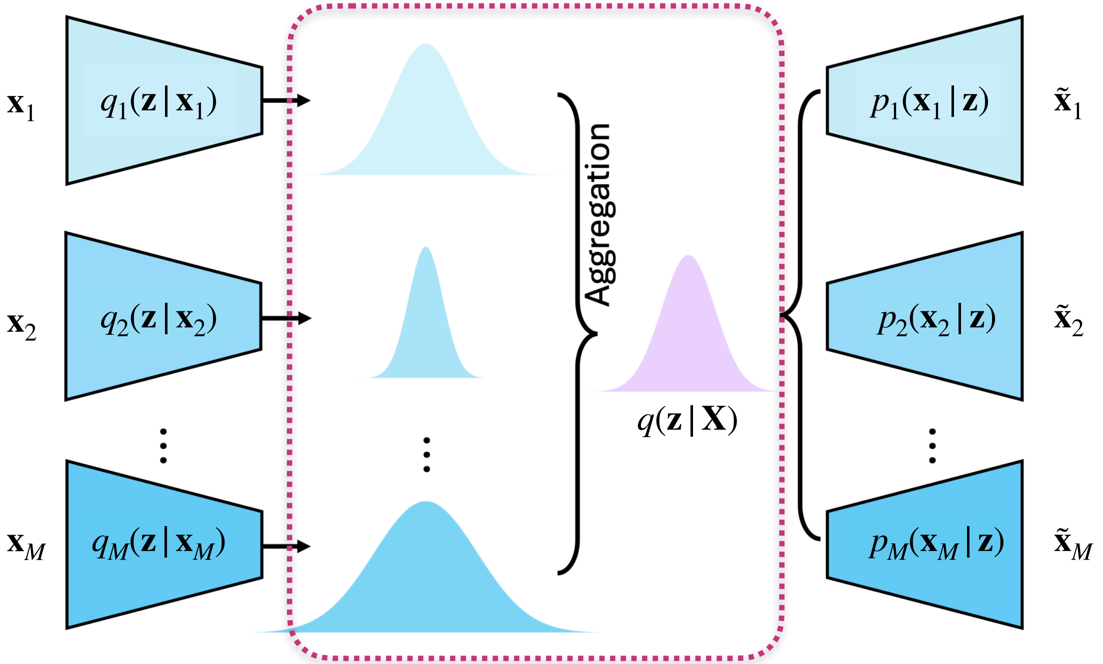


[1]

[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

Posteriors as **Gaussian** distributions



[1]

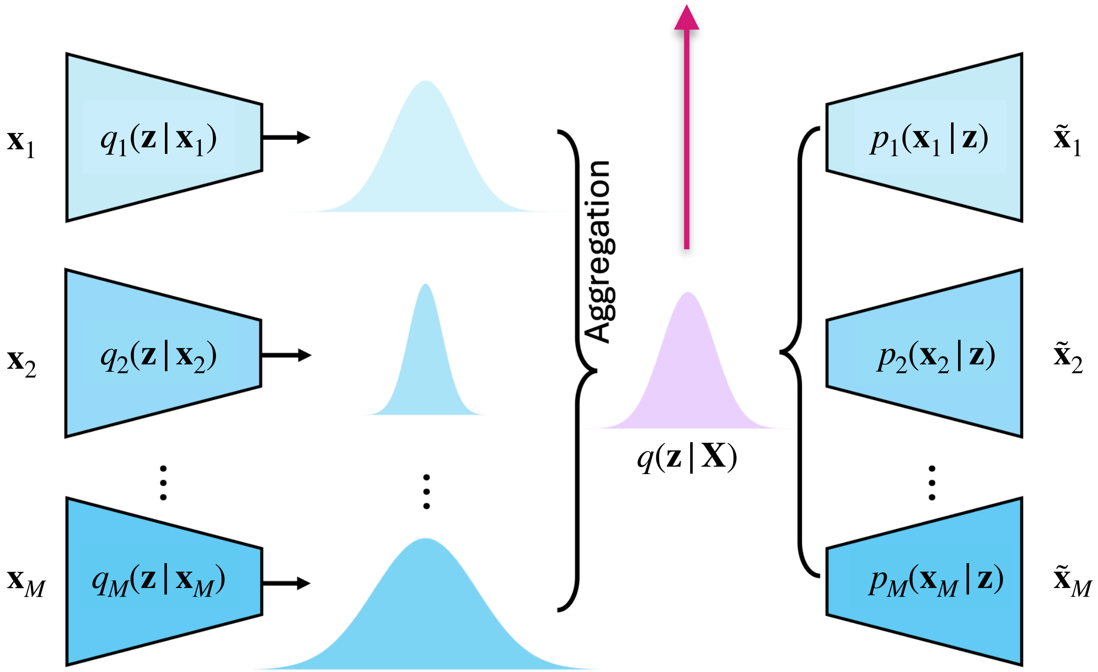
→ Multimodal VAEs learn **a joint posterior $q(\mathbf{z}|\mathbf{X})$** from M modalities using

M encoders $\{q_j(\mathbf{z}|\mathbf{x}_j)\}_{j=1}^{M}$ and M decoders $\{p_j(\mathbf{x}_j|\mathbf{z})\}_{j=1}^{M}$

[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

How to approximate the joint posterior $q(\mathbf{z} \mid \mathbf{X})$?



[1]

→ Multimodal VAEs learn **a joint posterior** $q(\mathbf{z} \mid \mathbf{X})$ from M modalities using M encoders $\{q_j(\mathbf{z} \mid \mathbf{x}_j)\}_{j=1}^{M}$ and M decoders $\{p_j(\mathbf{x}_j \mid \mathbf{z})\}_{j=1}^{M}$

[1] Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE):

Mixture of experts (MoE):

Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE):

$$q(\mathbf{z} \mid \mathbf{X}) = c \prod_{j=1}^{M} q_j(\mathbf{z} \mid \mathbf{x}_j)$$

Mixture of experts (MoE):

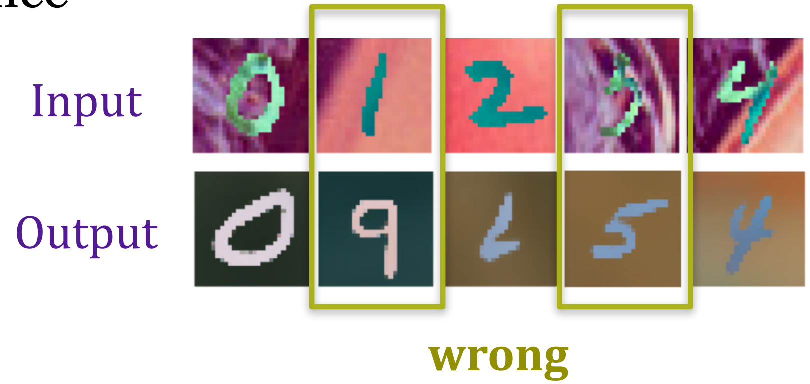
Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE): ✅ high quality, ❌ low coherence

$$q(\mathbf{z}\,|\,\mathbf{X}) = c\prod_{j=1}^{M} q_j(\mathbf{z}\,|\,\mathbf{x}_j)$$



Mixture of experts (MoE):

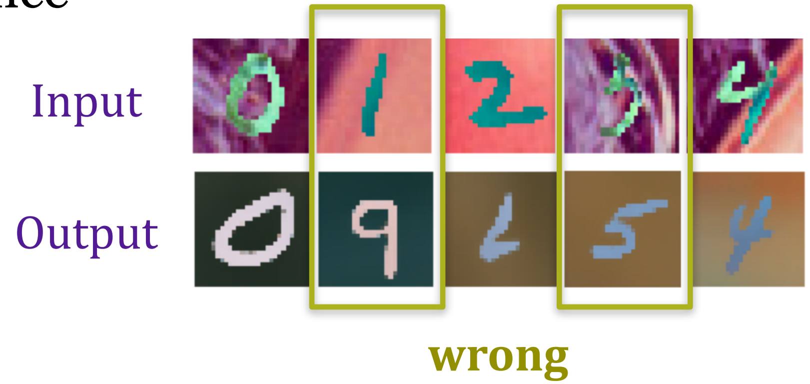
Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE): ✅ high quality, ❌ low coherence

$$q(\mathbf{z}\,|\,\mathbf{X}) = c\prod_{j=1}^{M} q_j(\mathbf{z}\,|\,\mathbf{x}_j)$$



Mixture of experts (MoE):

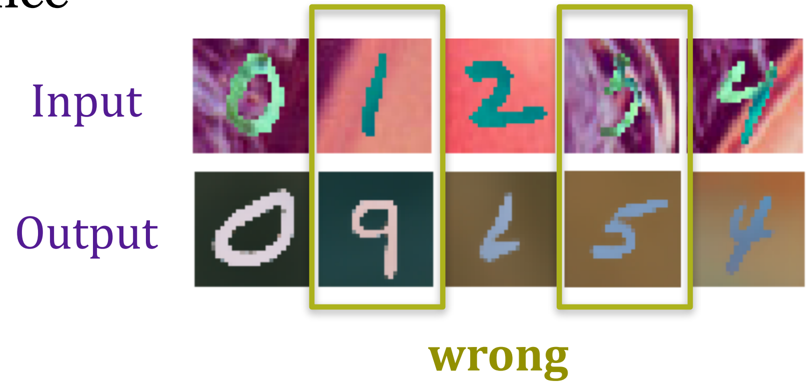
$$q(\mathbf{z}\,|\,\mathbf{X}) = \frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{z}\,|\,\mathbf{x}_j)$$

Motivation

Current Gaps

Product of experts (PoE): ✅ high quality, ❌ low coherence

$$q(\mathbf{z} \mid \mathbf{X}) = c \prod_{j=1}^{M} q_j(\mathbf{z} \mid \mathbf{x}_j)$$



Mixture of experts (MoE): ❌ low quality, ✅ high coherence

$$q(\mathbf{z} \mid \mathbf{X}) = \frac{1}{M} \sum_{j=1}^{M} q_j(\mathbf{z} \mid \mathbf{x}_j)$$

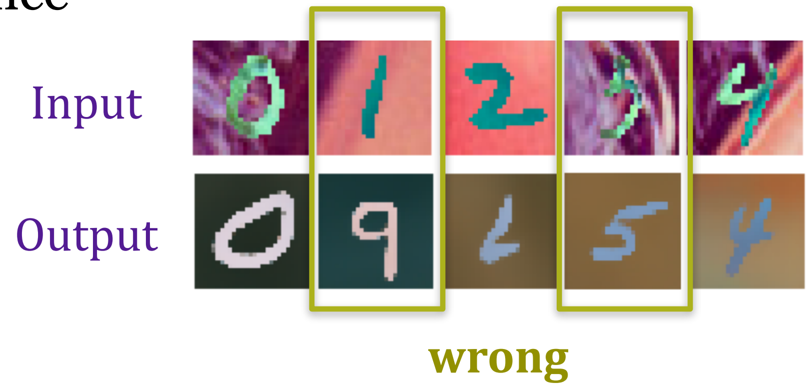
Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE): ✅ high quality, ❌ low coherence

$$q(\mathbf{z}\,|\,\mathbf{X}) = c\prod_{j=1}^{M} q_j(\mathbf{z}\,|\,\mathbf{x}_j)$$



Mixture of experts (MoE): ❌ low quality, ✅ high coherence

$$q(\mathbf{z}\,|\,\mathbf{X}) = \frac{1}{M}\sum_{j=1}^{M} q_j(\mathbf{z}\,|\,\mathbf{x}_j)$$

→ a trade-off between quality and coherence

Motivation

UNIVERSITÄT DES SAARLANDES

Current Gaps

Product of experts (PoE): ✅ high quality, ❌ low coherence

Mixture of experts (MoE): ❌ low quality, ✅ high coherence

→ a trade-off between quality and coherence

Research Question

Can we design a model that achieves

both high quality and high coherence while remaining efficient?

Multimodal aggregation as Probabilistic Opinion Pooling

Multimodal aggregation as Probabilistic Opinion Pooling

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Multimodal aggregation as Probabilistic Opinion Pooling

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

Multimodal aggregation as Probabilistic Opinion Pooling

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

$\alpha \to 0$

(reverse KL)

Product of experts (**PoE**)

✅ high quality, ❌ low coherence

Multimodal aggregation as Probabilistic Opinion Pooling

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

$\alpha \to 0$

(reverse KL)

Product of experts (**PoE**)

✅ high quality, ❌ low coherence

$\alpha = 1$

(forward KL)

Mixture of experts (**MoE**)

❌ low quality, ✅ high coherence

Multimodal aggregation as Probabilistic Opinion Pooling

UNIVERSITÄT DES SAARLANDES

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

$\alpha \to 0$

(reverse KL)

Product of experts (**PoE**)

✅ high quality, ❌ low coherence

Both measures are asymmetric and unbounded

$\alpha = 1$

(forward KL)

Mixture of experts (**MoE**)

❌ low quality, ✅ high coherence

Multimodal aggregation as Probabilistic Opinion Pooling

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

$\alpha \to 0$

(reverse KL)

Product of experts (**PoE**)

asymmetric & unbounded

✅ high quality, ❌ low coherence

$\alpha = 0.5$

unique symmetric

bounded

$\alpha = 1$

(forward KL)

Mixture of experts (**MoE**)

asymmetric & unbounded

❌ low quality, ✅ high coherence

Multimodal aggregation as Probabilistic Opinion Pooling

Universität des Saarlandes

Aggregating experts by minimizing the discrepancy between the individual and aggregated distributions.

Using the α-divergence yields the closed-form solution known as **Hölder pooling:**

$$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j (q_j(\mathbf{z}))^{\alpha}\right)^{1/\alpha}$$

$\alpha \to 0$

(reverse KL)

Product of experts (**PoE**)

asymmetric & unbounded

✅ high quality, ❌ low coherence

$\alpha = 0.5$

unique symmetric

bounded

✅ high quality ✅ high coherence ??

$\alpha = 1$

(forward KL)

Mixture of experts (**MoE**)

asymmetric & unbounded

❌ low quality, ✅ high coherence

Hellinger aggregation in Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Hölder pooling ($\alpha = 0.5$)

A mixture-of-Gaussians form

❌ Requires sub-sampling

Hellinger aggregation in Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Hölder pooling ($\alpha = 0.5$)

A mixture-of-Gaussians form

❌ Requires sub-sampling

Hellinger aggregation

a single-Gaussian approximation of Hölder pooling ($\alpha = 0.5$) via moment matching,

$$\tilde{\mu} = \int \mathbf{z} q(\mathbf{z}) d\mathbf{z}, \quad \tilde{\boldsymbol{\Sigma}} = \int \mathbf{z}\mathbf{z}^\top q(\mathbf{z}) d\mathbf{z} - \tilde{\mu}\tilde{\mu}^\top$$

Hellinger aggregation in Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Hölder pooling ($\alpha = 0.5$)

A mixture-of-Gaussians form

❌ Requires sub-sampling

Hellinger aggregation

a single-Gaussian approximation of Hölder pooling ($\alpha = 0.5$) via moment matching,

$$\tilde{\mu} = \int \mathbf{z} q(\mathbf{z}) d\mathbf{z}, \quad \tilde{\boldsymbol{\Sigma}} = \int \mathbf{z}\mathbf{z}^\top q(\mathbf{z}) d\mathbf{z} - \tilde{\mu}\tilde{\mu}^\top$$

✅ No sub-sampling

Hellinger aggregation in Multimodal VAEs

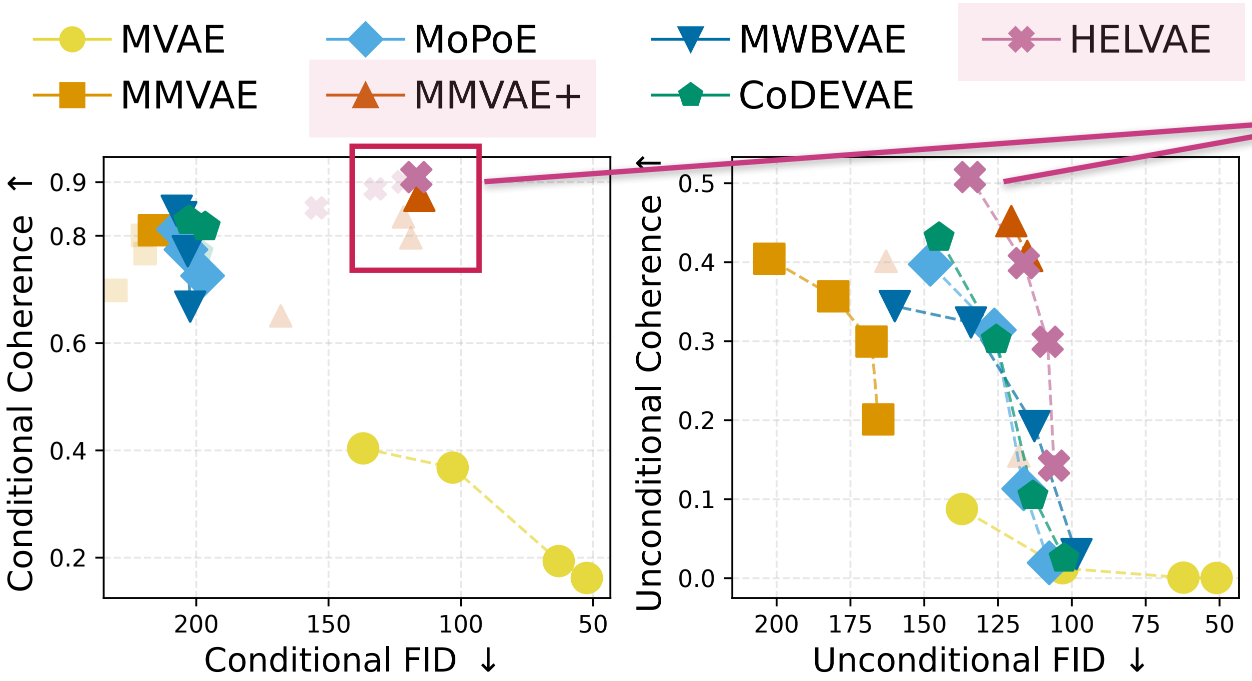
Hellinger aggregation

a single-Gaussian approximation of Hölder pooling ($\alpha = 0.5$) via moment matching,

$$\tilde{\mu} = \int \mathbf{z} q(\mathbf{z}) d\mathbf{z}, \quad \tilde{\Sigma} = \int \mathbf{z}\mathbf{z}^\top q(\mathbf{z}) d\mathbf{z} - \tilde{\mu}\tilde{\mu}^\top$$

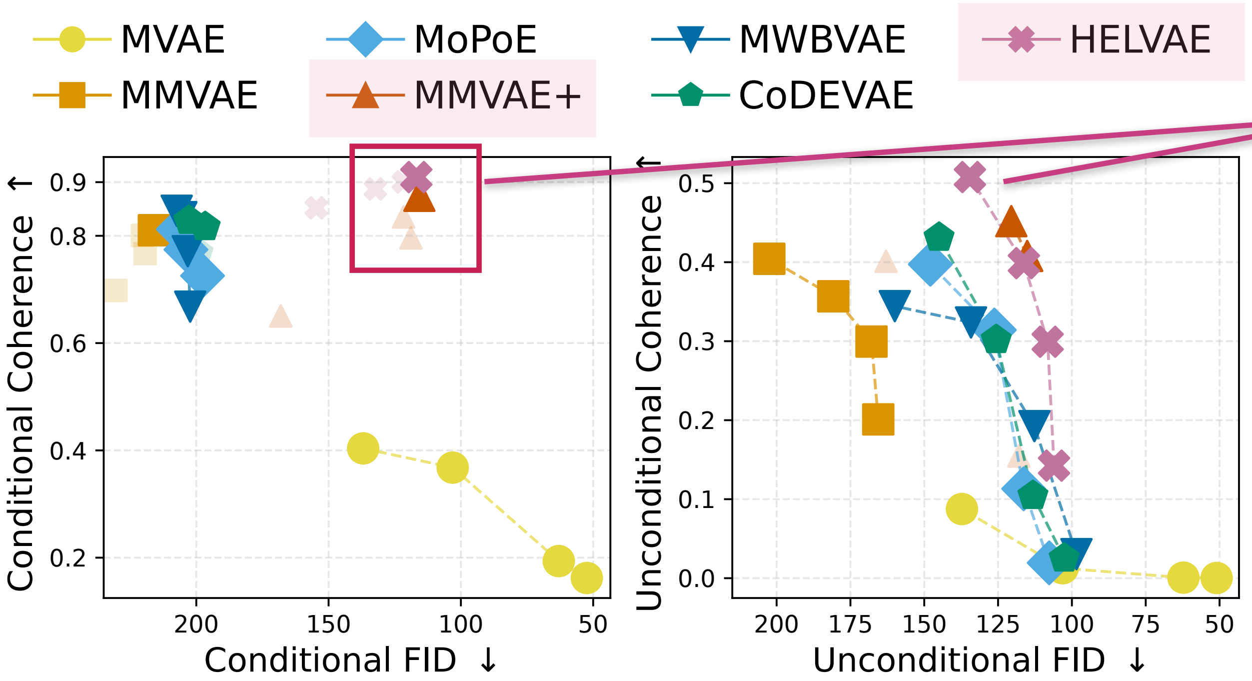
Hellinger aggregation defines a novel multimodal VAE, called **HELVAE**

Experimental Results



HELVAE lies close to the Pareto frontier, outperforming the SOTA **MMVAE+**

Experimental Results

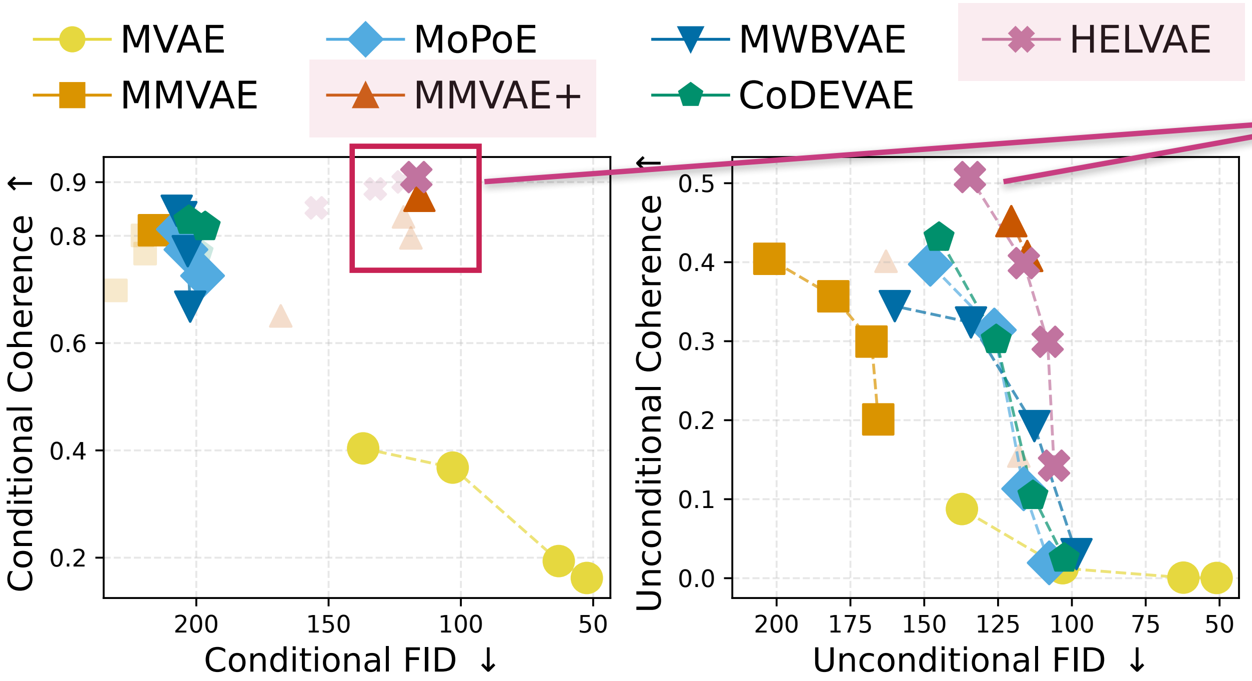


HELVAE lies close to the Pareto frontier, outperforming the SOTA **MMVAE+**

HELVAE	**MMVAE+**
0.0895 seconds/batch	0.2884 seconds/batch

Up to **3× faster** batch training

Experimental Results



HELVAE lies close to the Pareto frontier, outperforming the SOTA **MMVAE+**

HELVAE	**MMVAE+**
0.0895 seconds/batch	0.2884 seconds/batch

Up to **3× faster** batch training

HELVAE: ✅ high coherence ✅ high quality ✅ computationally efficient

→ an efficient yet effective model

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

erc European Research Council
Established by the European Commission

AISTATS

UNIVERSITÄT
DES
SAARLANDES

Paper

Code

Today at **Poster #186**
4pm - 7pm

🏫 Hiring Postdoc

Huyen Vo

Isabel Valera