Hölder++: Improving the Quality-Coherence Trade-off in Multimodal VAEs

Huyen Vo[1,2], **María Martínez-García**[1], Isabel Valera[1,2]

[1]Department of Computer Science, Saarland University, DE
[2]MPI-SWS, Saarland Informatics Campus, DE

Talk at BI, Department of Data Science and Analytics



European Research Council
Established by the European Commission



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS



UNIVERSITÄT
DES
SAARLANDES



ICML
International Conference
On Machine Learning

Outline

1. Background on Multimodal VAEs
2. **Hölder:** The first exact symmetric Hölder pooling as aggregation
3. **Hölder+:** Learning shared and private subspaces
4. **Hölder++:** Disentangling private and shared subspaces
5. Experimental Results
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

Outline

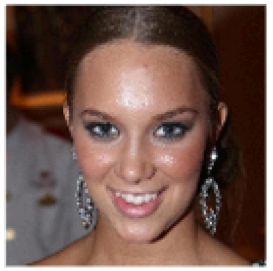
1. **Background on Multimodal VAEs**
2. **Hölder:** The first exact symmetric Hölder pooling as aggregation
3. **Hölder+:** Learning shared and private subspaces
4. **Hölder++:** Disentangling private and shared subspaces
5. Experimental Results
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

Why multimodal generative models?

UNIVERSITÄT DES SAARLANDES

- Real-world data often comes in **multiple modalities**: image, text, audio, …
- These modalities contain **shared information** about the same underlying concept, while each also carries **modality-specific information**.

Modality 1



Modality 2



Modality 3

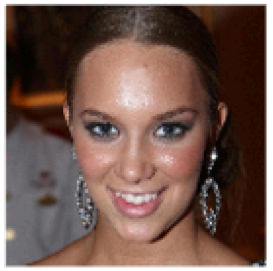
Brown_Hair
Heavy_Makeup
Mouth_Slightly_Open
Receding_Hairline
Smiling
Straight_Hair

Why multimodal generative models?

UNIVERSITÄT DES SAARLANDES

- Real-world data often comes in **multiple modalities**: image, text, audio, …
- These modalities contain **shared information** about the same underlying concept, while each also carries **modality-specific information**.

Modality 1



Modality 2



Modality 3

Brown_Hair
Heavy_Makeup
Mouth_Slightly_Open
Receding_Hairline
Smiling
Straight_Hair

- **Multimodal generative models** capture dependencies across modalities to support both **joint generation** and **cross-modal generation**.

What makes a good multimodal generative model?

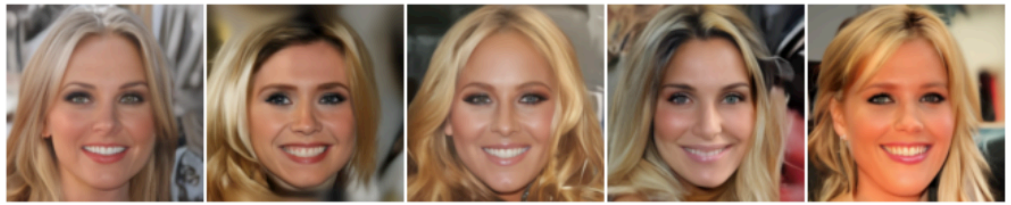
UNIVERSITÄT DES SAARLANDES

- **Generative quality:**
  - Generated samples from each modality should be **realistic and diverse**.
- **Generative coherence:**
  - Generated modalities should be **semantically consistent** with each other.

Input
(Mask + Attributes)



Cross-generation
(Image)



What makes a good multimodal generative model?

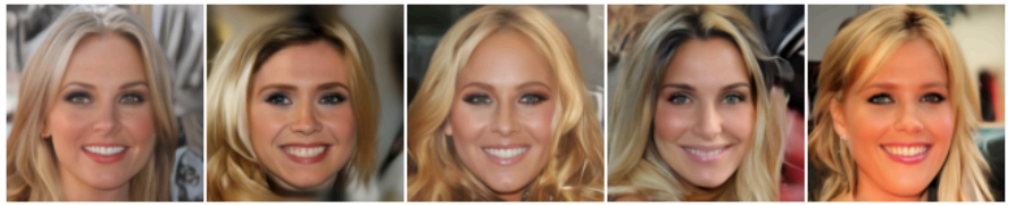
UNIVERSITÄT DES SAARLANDES

- **Generative quality:**
  - Generated samples from each modality should be **realistic and diverse**.
- **Generative coherence:**
  - Generated modalities should be **semantically consistent** with each other.

Input (Mask + Attributes)



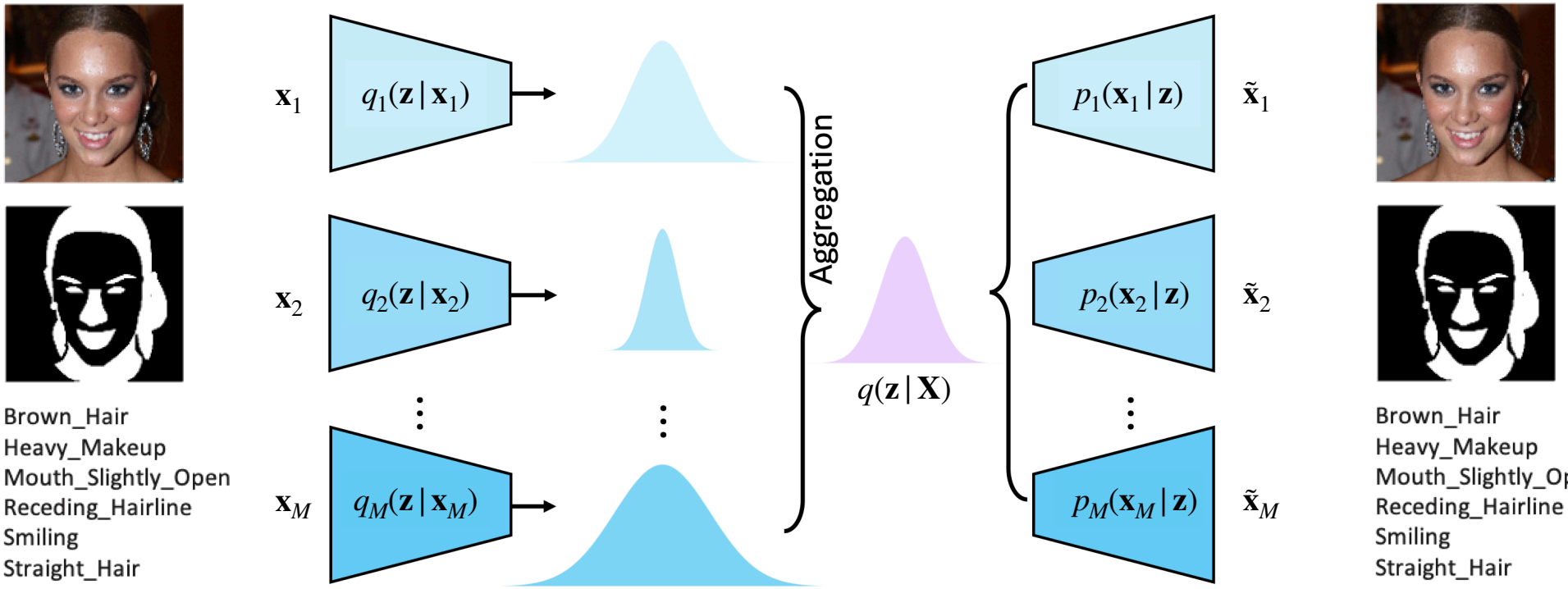
Cross-generation (Image)



Challenge: Improving coherence can reduce generation quality, and vice versa.

Multimodal VAEs

A **multimodal generative learning framework**

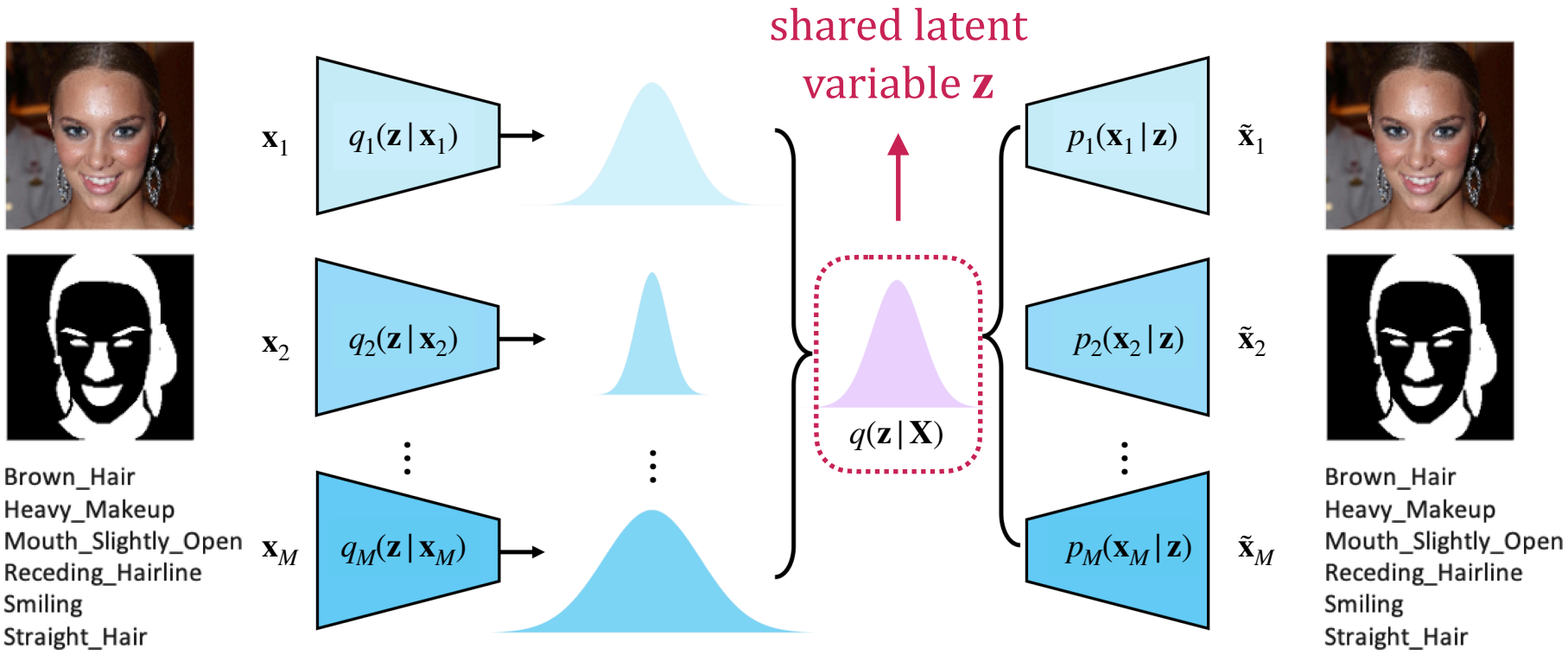


[1]

[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

We want **multimodal generative models** that can **(i) learn shared representations**,

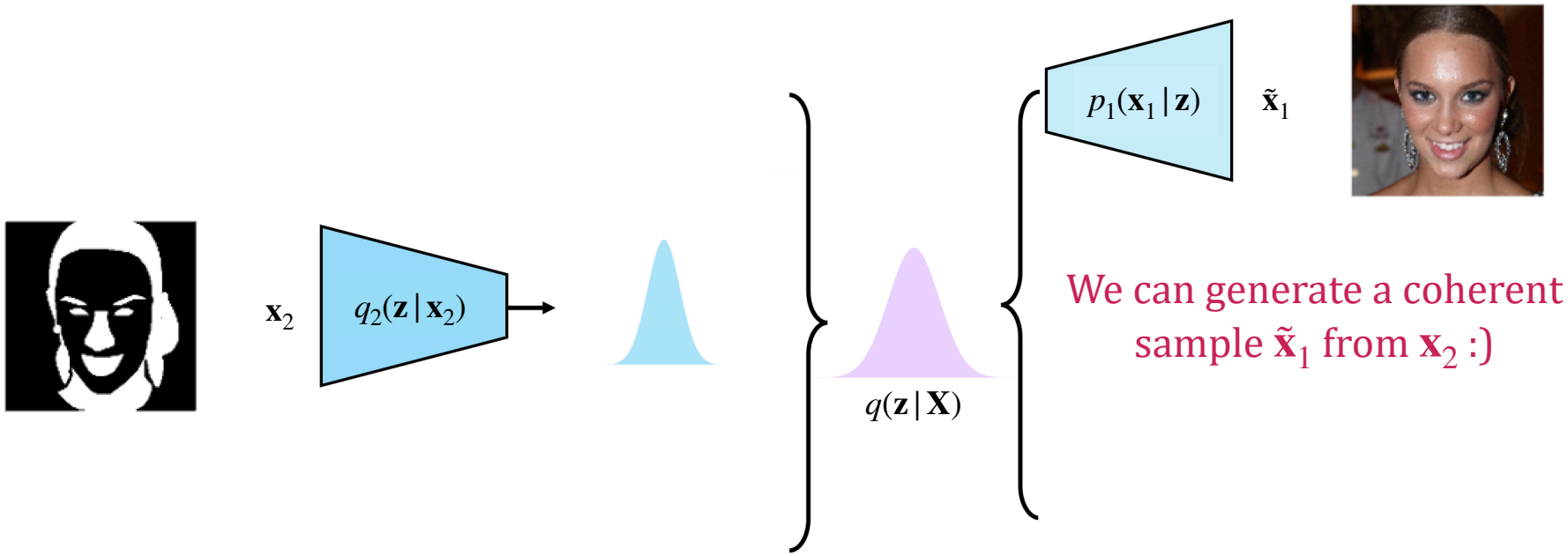


[1]

[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

We want **multimodal generative models** that can **(i) learn shared representations**, **(ii) support cross-modal generation**,

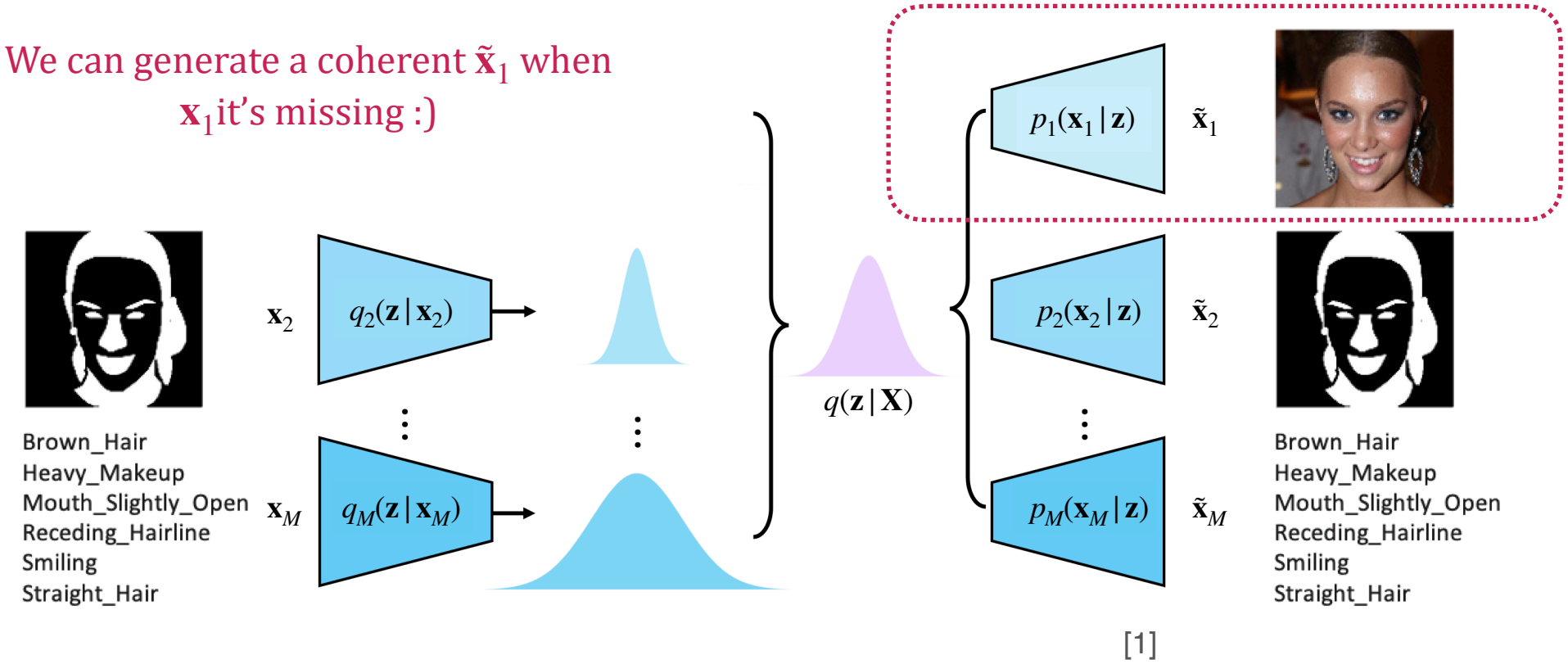


[1]

[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

We want **multimodal generative models** that can **(i) learn shared representations**, **(ii) support cross-modal generation**, and **(iii) handle missing modalities**.



[1]

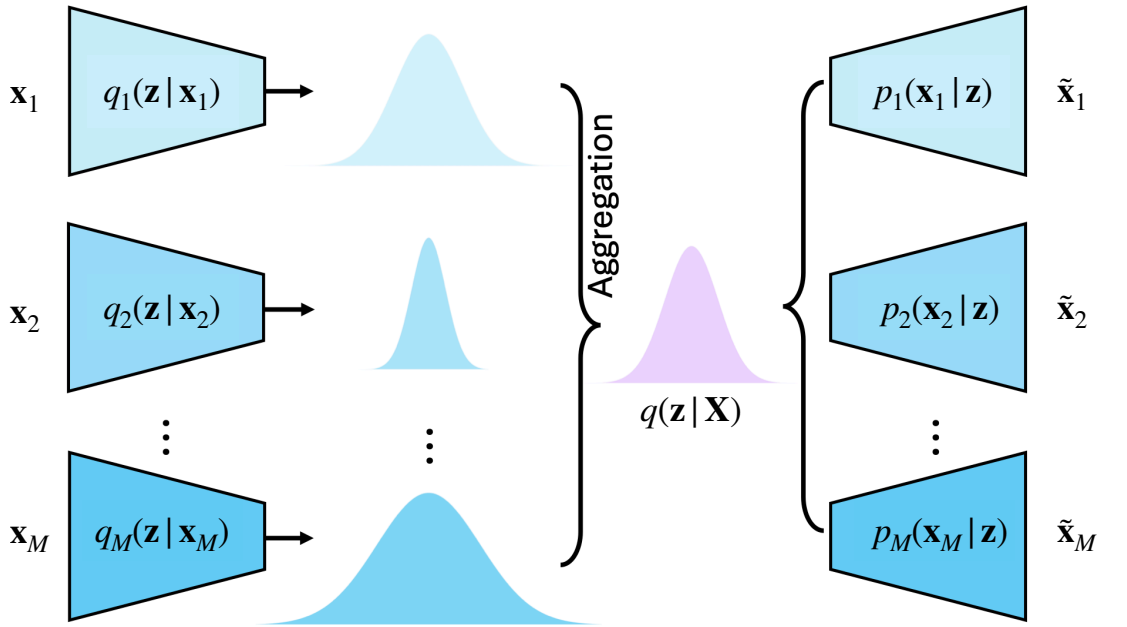
[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

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

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

Modality-specific variational posteriors are typically assumed to be **Gaussian**.



[1]

Training by maximizing the ELBO:

$$\mathscr{L}(\mathbf{x}_{1:M}) = \mathbb{E}_{q_\Phi(\mathbf{z}|\mathbf{X})}\left[\log \frac{p_\Theta(\mathbf{X},\mathbf{z})}{q_\Phi(\mathbf{z}\,|\,\mathbf{X})}\right]$$

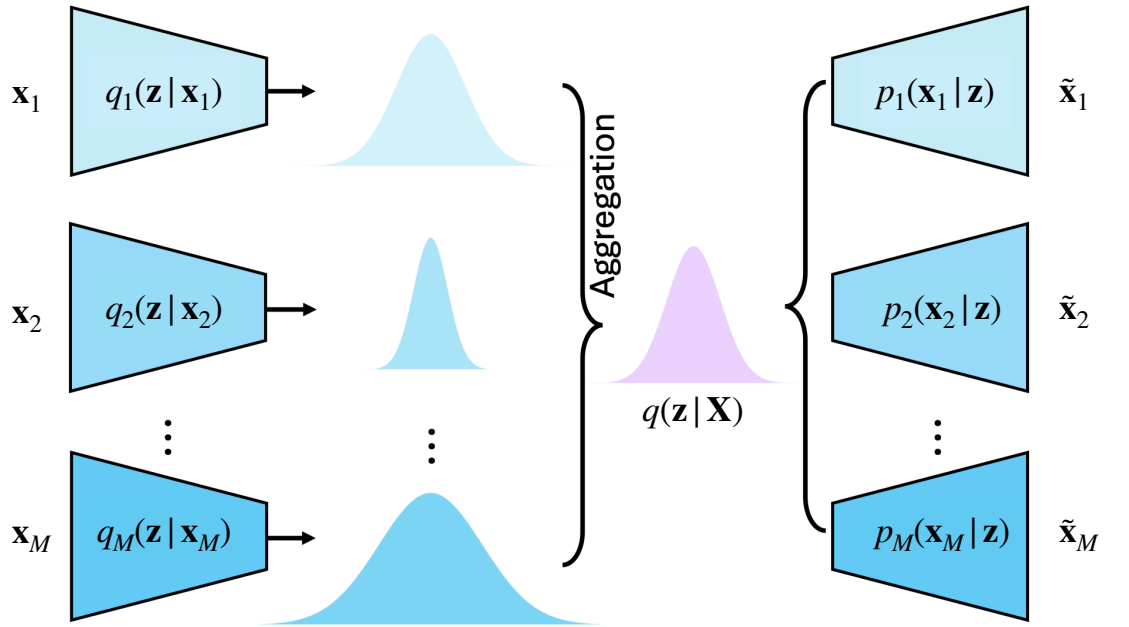
[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

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

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

Modality-specific variational posteriors are typically assumed to be **Gaussian**.



[1]

How can we approximate $q(\mathbf{z}|\mathbf{X})$?

[1] AAAI 2025. Qiu et al. Multimodal Variational Autoencoder: a Barycentric View.

Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Common approaches

Product of experts (PoE)

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

Mixture of experts (MoE)

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

Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Common approaches

Product of experts (PoE)

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

✅ high quality, ❌ low coherence



wrong digit!

Mixture of experts (MoE)

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

❌ low quality, ✅ high coherence



blurred, no diversity…

Multimodal VAEs

UNIVERSITÄT DES SAARLANDES

Our goal

Improve the quality-coherence trade-off to be able to generate **realistic**, **diverse** and **semantically consistent** samples across modalities.



✅ high quality, ✅ high coherence

Aggregation as Probabilistic Opinion Pooling

Aggregating experts by **minimizing the discrepancy** between the individual and aggregated distributions:

$$q(\mathbf{z}) = \arg\min_{\varphi \in \mathscr{P}} \sum_{j=1}^{M} \lambda_j \mathscr{D}_\alpha(q_j \| \varphi)$$

- $q(\mathbf{z})$: aggregated expert (joint) posterior
- $\mathscr{D}_\alpha$: chosen measure (α-**divergence**)
- $\{q_j(\mathbf{z})\}_{j=1}^{M}$: individual expert posteriors

[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

Aggregation as Probabilistic Opinion Pooling

Aggregating experts by **minimizing the discrepancy** between the individual and aggregated distributions:

$$q(\mathbf{z}) = \arg\min_{\varphi \in \mathcal{P}} \sum_{j=1}^{M} \lambda_j \mathcal{D}_\alpha(q_j \| \varphi)$$

- $q(\mathbf{z})$: aggregated expert (joint) posterior
- $\mathcal{D}_\alpha$: chosen measure (α-**divergence**)
- $\{q_j(\mathbf{z})\}_{j=1}^{M}$: individual expert posteriors

Closed-form solution as **Hölder pooling**:

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

[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

Aggregation as Probabilistic Opinion Pooling

UNIVERSITÄT DES SAARLANDES

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 and unbounded
- ✅ high quality
- ❌ low coherence

$\alpha = 1$

Forward KL
Mixture of Experts (MoE)

- Asymmetric and unbounded
- ❌ low quality
- ✅ high coherence

[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

Aggregation as Probabilistic Opinion Pooling

UNIVERSITÄT DES SAARLANDES

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 and unbounded
- ✅ high quality
- ❌ low coherence

$\alpha = 0.5$

Symmetric divergence

- Symmetric and bounded
- ✅ high quality?
- ✅ high coherence?

$\alpha = 1$

Forward KL
Mixture of Experts (MoE)

- Asymmetric and unbounded
- ❌ low quality
- ✅ high coherence

[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

HELVAE

Approximation of symmetric Hölder pooling

HELVAE [2] approximates the symmetric Hölder pooling via **moment matching.**

- **Hölder pooling:** $$q(\mathbf{z}) = c\left(\sum_{j=1}^{M} \lambda_j \sqrt{q_j(\mathbf{z})}\right)^2 = c\left(\sum_{j=1}^{M} \lambda_j^2 q_j(\mathbf{z}) + 2\sum_{i=1}^{M}\sum_{j>i}^{M} \lambda_i \lambda_j \sqrt{q_i(\mathbf{z}) q_j(\mathbf{z})}\right)$$

- **Moment matching approximation:** $$q(\mathbf{z}) \sim \mathcal{N}(\tilde{\mu}, \tilde{\mathbf{\Sigma}})$$

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

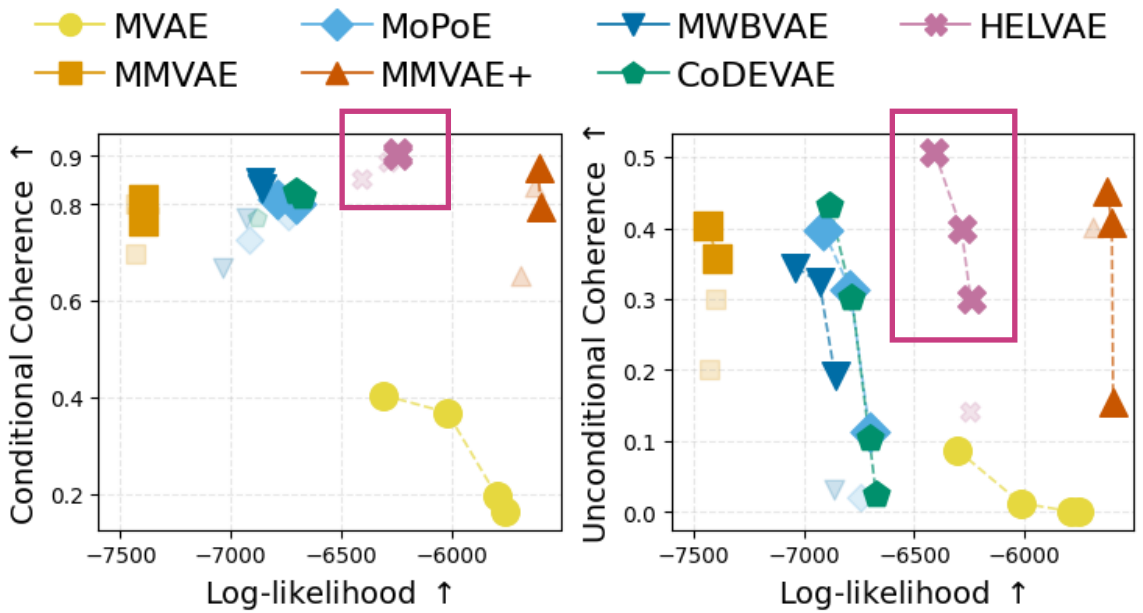
[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

HELVAE

Limitations observed during empirical evaluation

(1) Effect of the approximation: HELVAE

shows lower log-likelihood than **MMVAE+** (SOTA).

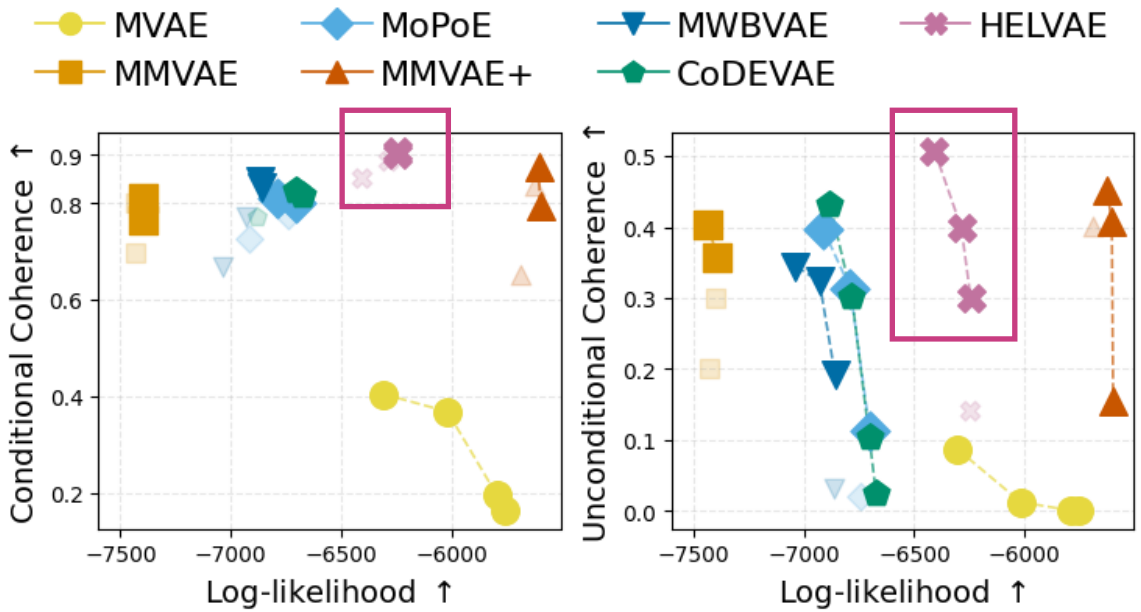


[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

HELVAE

Limitations observed during empirical evaluation

(1) Effect of the approximation: HELVAE shows lower log-likelihood than **MMVAE+** (SOTA).



(2) Single-shared representation: HELVAE shows low diversity compared to **MMVAE+** (SOTA).



(d) MMVAE+ (g) HELVAE

[2] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger Multimodal Variational Autoencoders.

Outline

1. Background on Multimodal VAEs
2. **Hölder: The first exact symmetric Hölder pooling as aggregation**
3. Hölder+: Learning shared and private subspaces
4. Hölder++: Disentangling private and shared subspaces
5. Experimental Results
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

Hölder: exact (symmetric) Hölder pooling as aggregation

We provide **the first exact implementation** of symmetric Hölder pooling as aggregation in the multimodal VAE framework.

- **Hölder:** $q(\mathbf{z}|\mathbf{X}) = \sum_{j=1}^{M} \pi_j \underbrace{q_{\phi_{\mathbf{z}_j}}(\mathbf{z}|\mathbf{x}_j)}_{\text{unimodal component}} + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \underbrace{q_{ij}^{(1/2)}(\mathbf{z}|\mathbf{x}_i, \mathbf{x}_j)}_{\text{pairwise component}},$

Hölder: exact (symmetric) Hölder pooling as aggregation

UNIVERSITÄT DES SAARLANDES

We provide **the first exact implementation** of symmetric Hölder pooling as aggregation in the multimodal VAE framework.

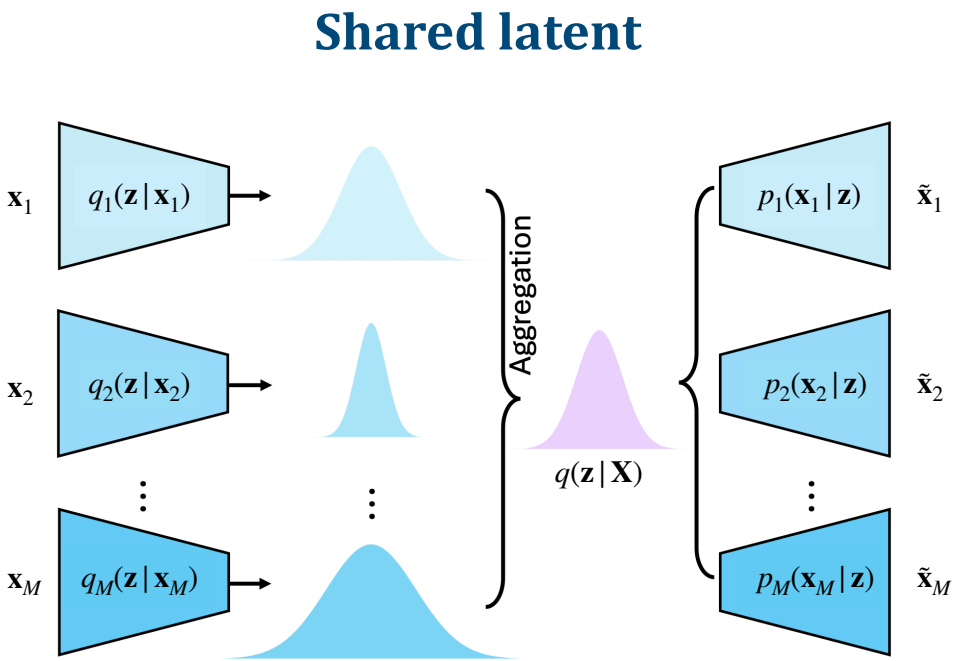
- **Hölder:** $q(\mathbf{z}\,|\,\mathbf{X}) = \sum_{j=1}^{M} \pi_j \underbrace{q_{\phi_{\mathbf{z}_j}}(\mathbf{z}\,|\,\mathbf{x}_j)}_{\text{unimodal component}} + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \underbrace{q_{ij}^{(1/2)}(\mathbf{z}\,|\,\mathbf{x}_i, \mathbf{x}_j)}_{\text{pairwise component}},$

ELBO

$$\mathscr{L}_{\text{Hölder}}(\mathbf{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{q_{\phi_j}(\mathbf{z}|\mathbf{x}_j)}\left[\log \frac{p_\Theta(\mathbf{X},\mathbf{z})}{q_\Phi(\mathbf{z}\,|\,\mathbf{X})}\right] + \sum_{i=1}^{M}\sum_{j>i}^{M} \pi_{ij} \mathbb{E}_{q_{ij}^{(1/2)}(\mathbf{z}|\mathbf{x}_i,\mathbf{x}_j)}\left[\log \frac{p_\Theta(\mathbf{X},\mathbf{z})}{q_\Phi(\mathbf{z}\,|\,\mathbf{X})}\right],$$

Hölder: exact (symmetric) Hölder pooling as aggregation

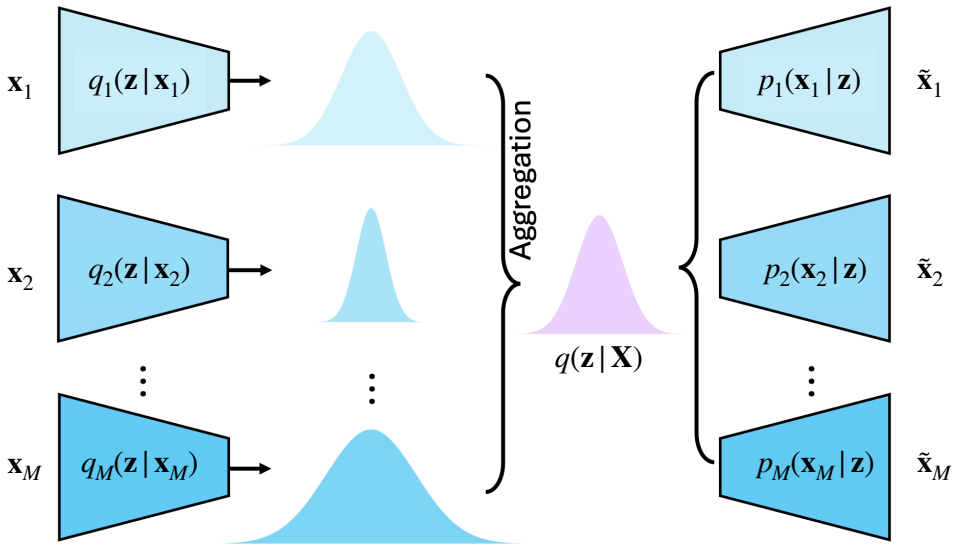
Both HELVAE and Hölder show **limited sample diversity** because they consider a **single shared representation across modalities**



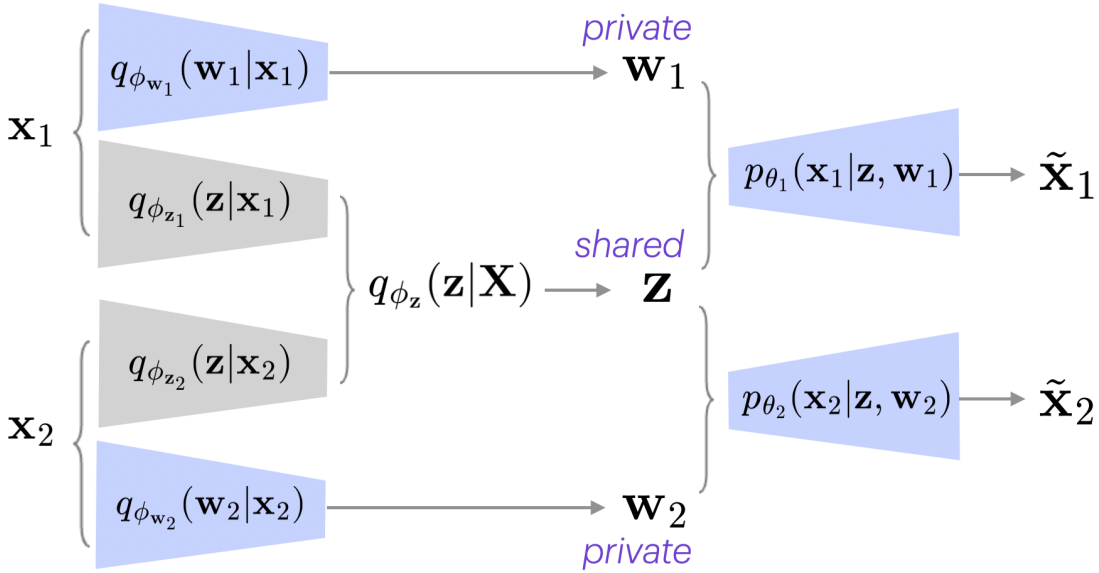
Hölder: exact (symmetric) Hölder pooling as aggregation

Both HELVAE and Hölder show **limited sample diversity** because they consider a **single shared representation across modalities** → **factorize** the latent space into **shared and private latents.**

Shared latent



Private + shared latents



Outline

1. Background on Multimodal VAEs
2. Hölder: The first exact symmetric Hölder pooling as aggregation
3. **Hölder+: Learning shared and private subspaces**
4. Hölder++: Disentangling private and shared subspaces
5. Experimental Results
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

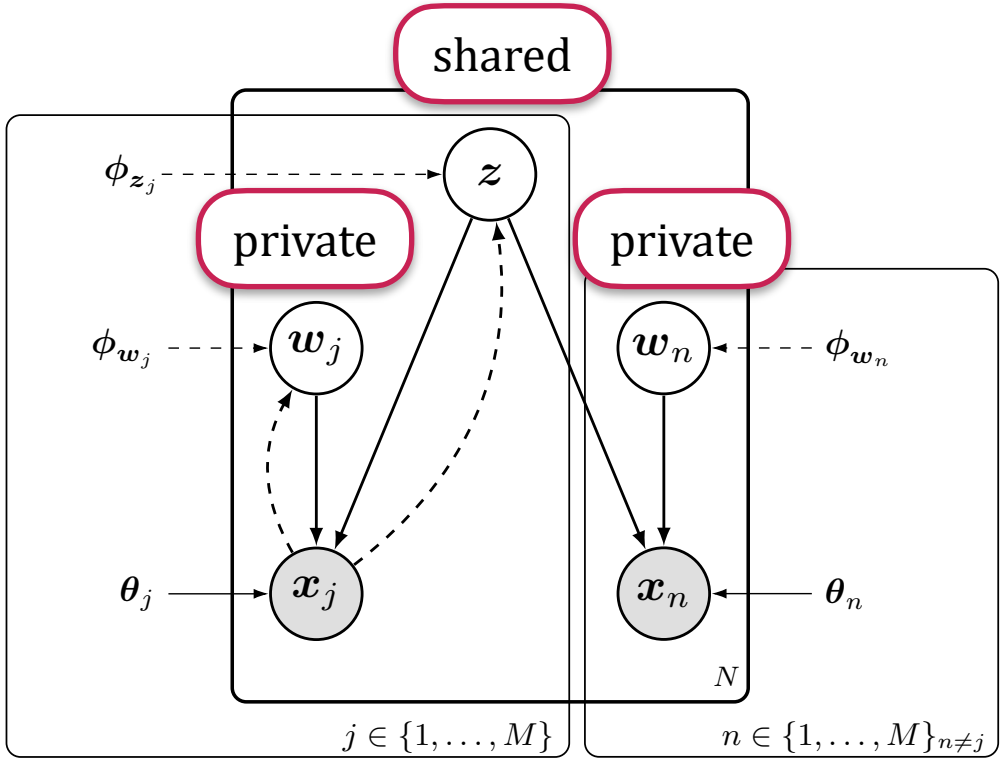
Hölder+: learning shared and private subspaces

We **factorize** the latent space into

- a shared latent $\mathbf{z}$, and
- M private latents $\mathbf{W} = \{\mathbf{w}_j\}_{j=1}^{M}$.

Generation: $p_\Theta(\mathbf{X}, \mathbf{z}, \mathbf{W}) = p(\mathbf{z}) \prod_{j=1}^{M} p_{\theta_j}(\mathbf{x}_j \mid \mathbf{z}, \mathbf{w}_j) p(\mathbf{w}_j)$

Inference: $q_\Phi(\mathbf{z}, \mathbf{W} \mid \mathbf{X}) = q_{\phi_z}(\mathbf{z} \mid \mathbf{X}) \prod_{j=1}^{M} q_{\phi_j}(\mathbf{w}_j \mid \mathbf{x}_j)$



Hölder+: learning shared and private subspaces

Potential issue: shortcut learning [3].

The model captures in $\mathbf{w}_n$ all the information necessary for the reconstruction of modality n, **ignoring the shared latent z**. This is, **the model learns a shortcut!**

Unimodal term in the vanilla factorized ELBO:

$$\mathscr{L}(\mathbf{x}_{1:M}) = \sum_{j=1}^{M} \pi_j \mathbb{E}_{\substack{q_{\phi_{\mathbf{z}_j}}(\mathbf{z}\,|\,\mathbf{x}_j) \\ q_{\phi_{\mathbf{w}_j}}(\mathbf{w}_j\,|\,\mathbf{x}_j) \\ \{\mathbf{w}_n \sim q_{\phi_{\mathbf{w}_n}}(\mathbf{w}_n\,|\,\mathbf{x}_n)\}_{n\neq j}}} \log \left(\frac{p_{\theta_j}(\mathbf{x}_j\,|\,\mathbf{z}, \mathbf{w}_j)p(\mathbf{z})p(\mathbf{w}_j)}{q_{\phi_{\mathbf{z}}}(\mathbf{z}\,|\,\mathbf{X})q_{\phi_{\mathbf{w}_j}}(\mathbf{w}_j\,|\,\mathbf{x}_j)} \prod_{n\neq j} p_{\theta_n}(\mathbf{x}_n\,|\,\mathbf{z}, \mathbf{w}_n) \right)$$

cross-reconstruction term

[3] ICLR 2023. Emanuele et al. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises.

Hölder+: learning shared and private subspaces

UNIVERSITÄT DES SAARLANDES

Potential issue: shortcut learning [3].

The model captures in $\mathbf{w}_n$ all the information necessary for the reconstruction of modality n, **ignoring the shared latent z**. This is, **the model learns a shortcut!**

But we cannot obtain high coherence if the shared latent is ignored!

We need informative shared representations!

[3] ICLR 2023. Emanuele et al. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises.

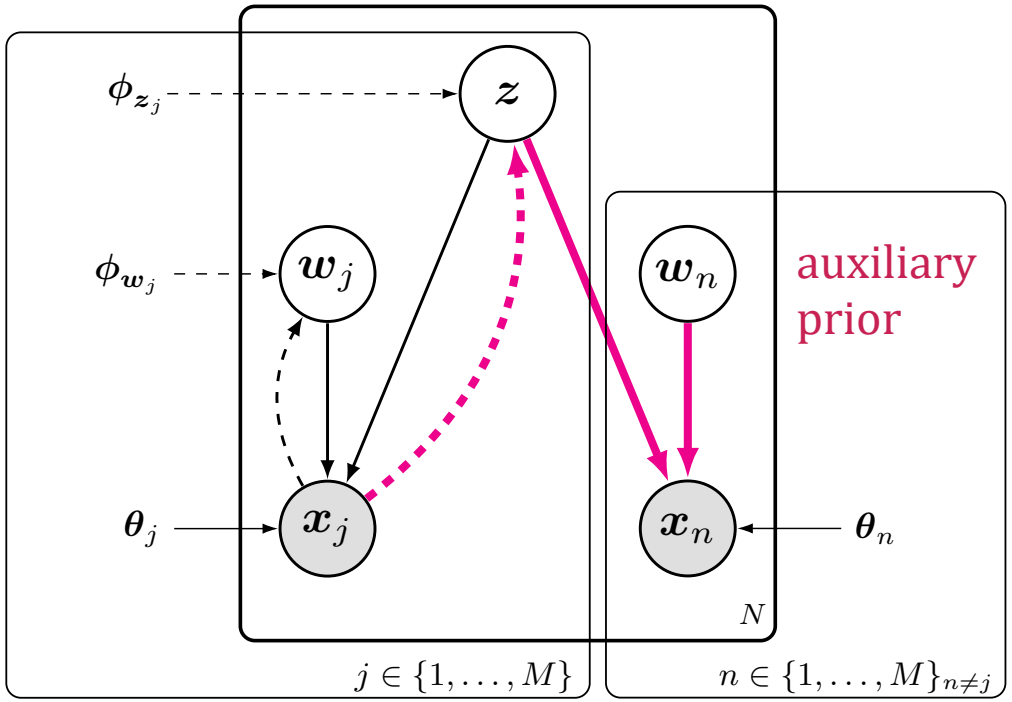
Hölder+: learning shared and private subspaces

Shortcut prevention [3]:

When $\mathbf{z} \sim q_{\phi_j}(\mathbf{z} \mid \mathbf{x}_j)$, each $\mathbf{w}_n$ is sampled as

$$\mathbf{w}_n \sim \begin{cases} q_{\phi_{\mathbf{w}_n}}(\mathbf{w}_n \mid \mathbf{x}_n), & n = j, \quad \text{self-term} \\ r_n(\mathbf{w}_n), & n \neq j \quad \text{cross-term} \end{cases}$$

During training, the private latent for cross-generation is sampled from an **uninformative auxiliary prior.**



[3] ICLR 2023. Emanuele et al. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises.

Outline

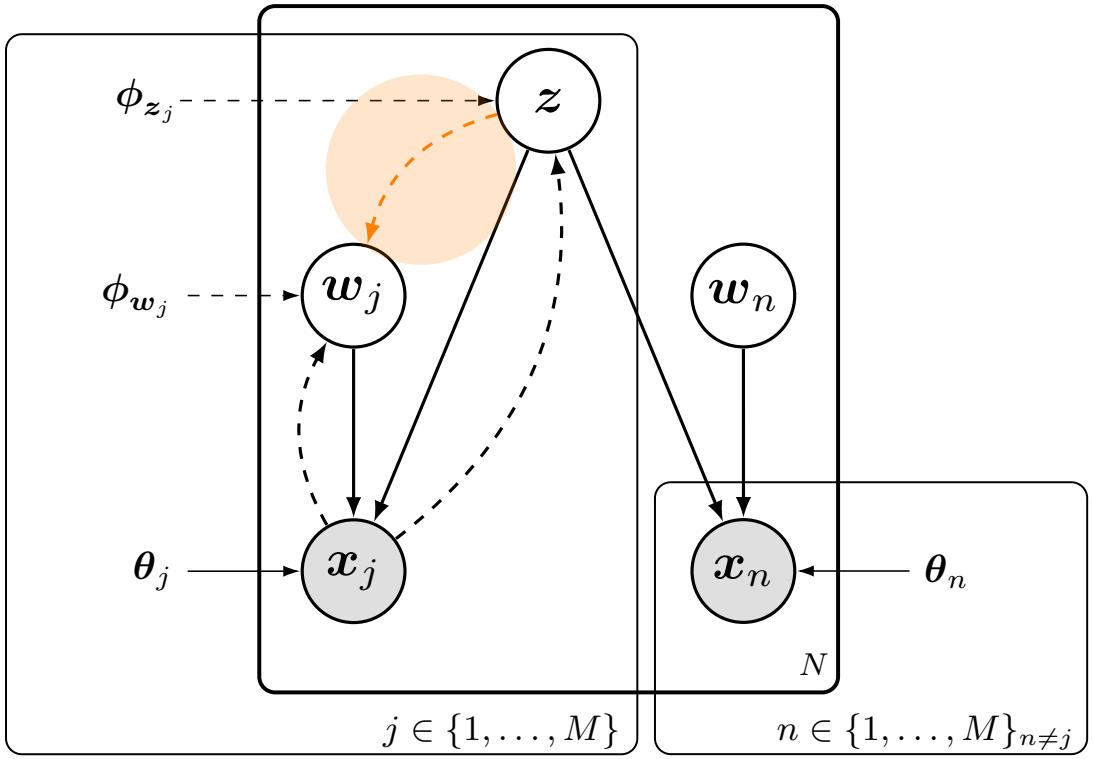
1. Background on Multimodal VAEs
2. Hölder: The first exact symmetric Hölder pooling as aggregation
3. Hölder+: Learning shared and private subspaces
4. **Hölder++: Disentangling private and shared subspaces**
5. Experimental Results
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

Hölder++: disentangling private-shared subspaces

We propose a variational factorization that **encourages disentanglement *by design*.**

Joint posterior:

$$q_\Phi(\mathbf{z}, \mathbf{W} \mid \mathbf{X}) = q_{\Phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{X}) \prod_{j=1}^{M} q_{\phi_{\mathbf{w}_j}}(\mathbf{w}_j \mid \mathbf{x}_j, \mathbf{z})$$



Hölder++: disentangling private-shared subspaces

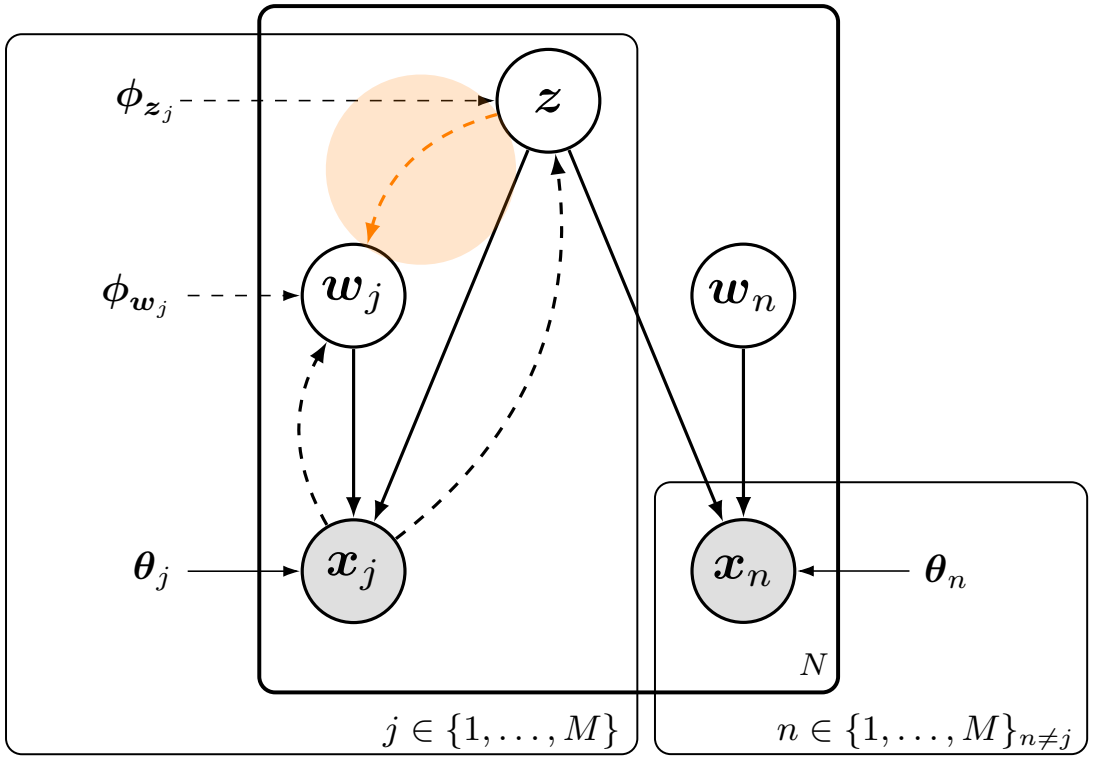
We propose a variational factorization that **encourages disentanglement *by design*.**

Joint posterior:

$$q_\Phi(\mathbf{z}, \mathbf{W} \mid \mathbf{X}) = q_{\Phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{X}) \prod_{j=1}^{M} q_{\phi_{\mathbf{w}_j}}(\mathbf{w}_j \mid \mathbf{x}_j, \mathbf{z})$$

Intuition:

1. Infer the **shared representation z** (semantics).
2. Infer each **modality-specific latent $\mathbf{w}_j$ conditioned** on both the input $\mathbf{x}_j$ and the shared latent **z**.



Hölder++: disentangling private-shared subspaces

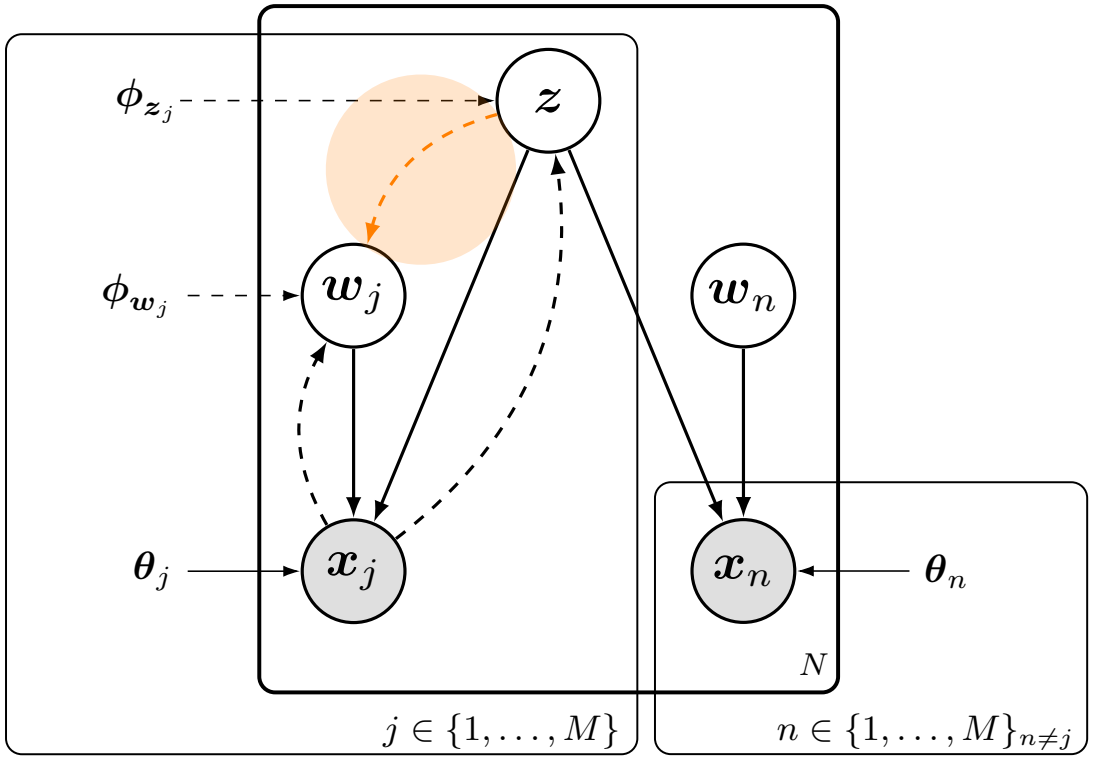
We propose a variational factorization that **encourages disentanglement *by design*.**

Joint posterior:

$$q_\Phi(\mathbf{z}, \mathbf{W} \mid \mathbf{X}) = q_{\Phi_\mathbf{z}}(\mathbf{z} \mid \mathbf{X}) \prod_{j=1}^{M} q_{\phi_{\mathbf{w}_j}}(\mathbf{w}_j \mid \mathbf{x}_j, \mathbf{z})$$

Intuition:

The private latents $\{\mathbf{w}_j\}_{j=1}^{M}$ model **modality-specific information** that is **not already explained** by the shared latent **z**.



Outline

1. Background on Multimodal VAEs
2. Hölder: The first exact symmetric Hölder pooling as aggregation
3. Hölder+: Learning shared and private subspaces
4. Hölder++: Disentangling private and shared subspaces
5. **Experimental Results**
    a. Generative quality and coherence trade-off
    b. Disentanglement of shared and private subspaces

Experimental Setup

Datasets

- PolyMNIST (5 modalities)



- MNIST-SVHN (2 modalities)



- CUBICC (2 modalities)



- CelebAMask-HQ (3 modalities)



Experimental Setup

Metrics

Generative quality and coherence

- **Generative quality (FID)**: measures the realism and diversity of generated samples.
- **Cross-modal coherence**: captures semantic consistency across modalities.
- **Log-likelihood**: assess how well the model fits the observed data distribution.

Experimental Setup

Metrics

We propose three new disentanglement metrics

- **w content accuracy (↓):** fix $\mathbf{w}$, sample multiple $\mathbf{z} \sim p(\mathbf{z})$, decode, classify and measure accuracy.

- **z content accuracy (↑)** and **z content stability (↑):** fix $\mathbf{z}$, sample multiple $\mathbf{w} \sim p(\mathbf{w})$, decode, classify and measure accuracy and invariant content (given variations in $\mathbf{w}$).

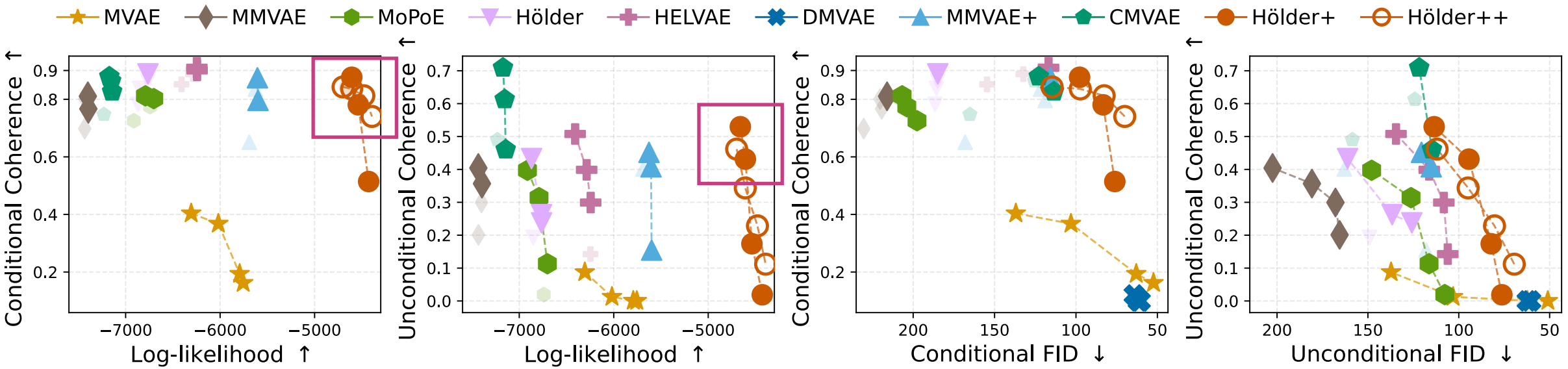


5 MNIST samples generated from SVHN by varying $\mathbf{w}$ and fixing $\mathbf{z}$.

Experimental Results

Generative quality and coherence trade-off

Hölder+/++ achieves the highest log-likelihood :)

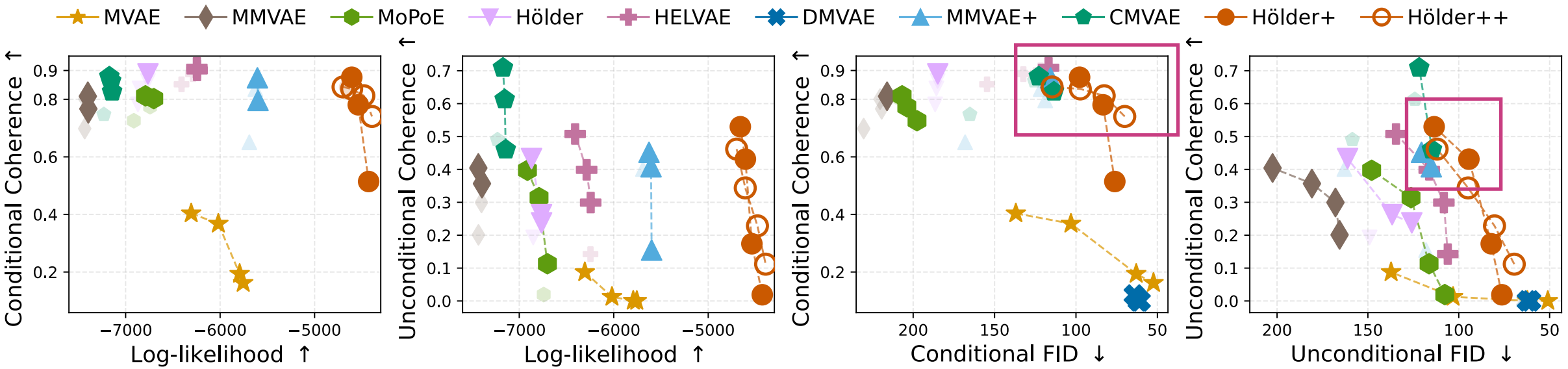


Optimal region: upper-right for all plots

Experimental Results

Generative quality and coherence trade-off

Hölder+/++ lies close to the Pareto frontier, outperforming the SOTA **MMVAE+!**

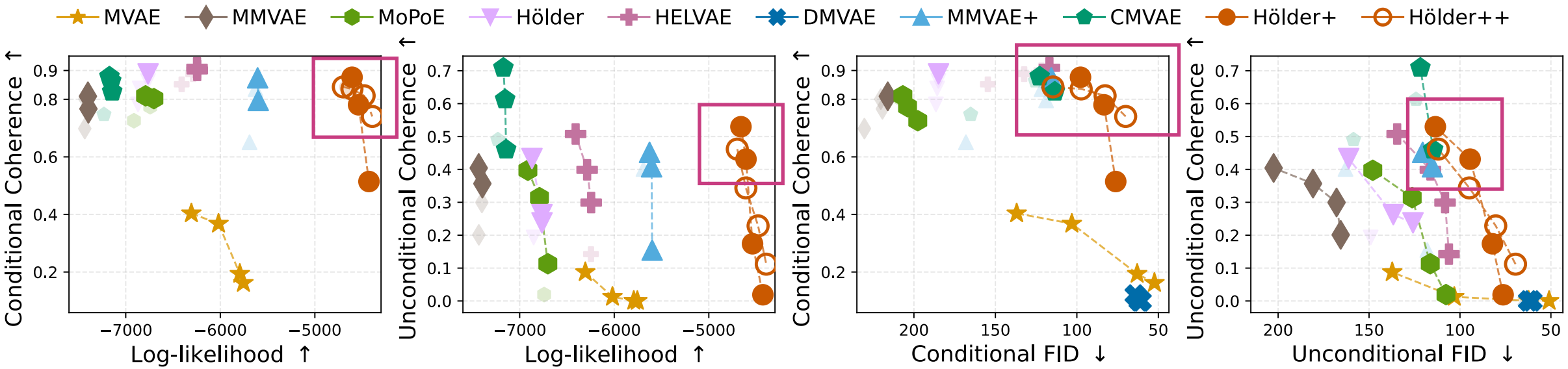


Optimal region: upper-right for all plots

Experimental Results

Generative quality and coherence trade-off

Main result: Hölder+/++ achieve **the best trade-off** between generative coherence and quality (FID).



Optimal region: upper-right for all plots

Experimental Results

Disentanglement of shared and private subspaces

Table 6. Effect of hierarchical inference on CUBICC. The best and second-best results are marked bold and underlined, respectively.

Method	w content accuracy ↓	z content stability ↑	z content accuracy ↑
MMVAE+	0.138 ± 0.013	0.846 ± 0.166	0.624 ± 0.049
MMVAE++	0.133 ± 0.002	**0.960 ± 0.008**	0.640 ± 0.010
Hölder+	0.195 ± 0.048	0.442 ± 0.087	0.443 ± 0.051
Hölder++	0.133 ± 0.004	0.870 ± 0.097	0.619 ± 0.021
CMVAE	0.136 ± 0.011	0.893 ± 0.145	0.641 ± 0.060
CMVAE++	**0.130 ± 0.003**	<u>0.959 ± 0.008</u>	**0.651 ± 0.011**
CHölder+	0.142 ± 0.006	0.587 ± 0.045	0.535 ± 0.030
CHölder++	<u>0.132 ± 0.003</u>	0.914 ± 0.066	<u>0.642 ± 0.014</u>

Main result: Hölder++ (hierarchical inference) **improves disentanglement** between shared and private latents over **Hölder+.**

Experimental Results

Computational cost

Our method requires only about $1.8\times$ the time of MMVAE-based baselines.

→ **We do not observe a quadratic increase in training time as M grows.**

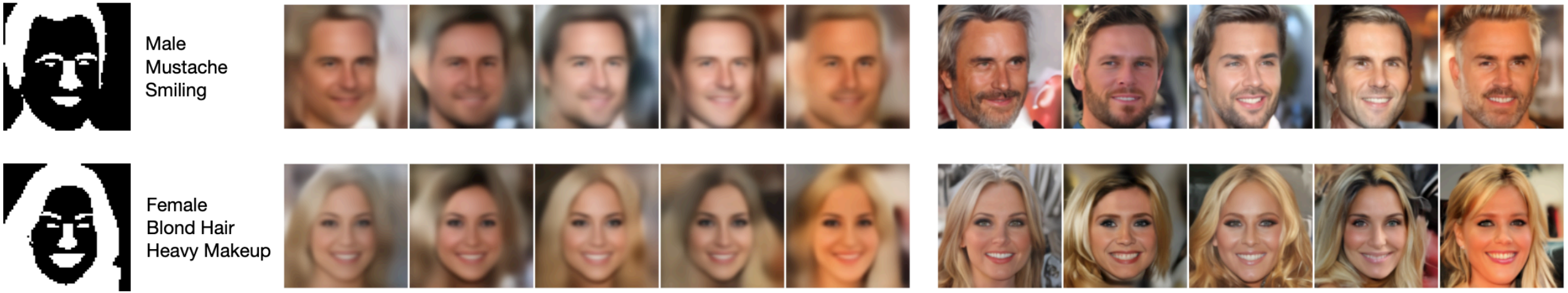
Table 8. Training batch time (s) across datasets. We group models by latent structure: shared-only vs. shared-private.

Model	MNIST-SVHN ($M = 2$)	CelebAMask-HQ ($M = 3$)	PolyMNIST ($M = 5$)
MVAE	0.0482	0.1442	0.1887
MoPoE	0.0321	0.0747	0.1216
HELVAE	0.0303	0.0727	0.0895
DMVAE	0.0537	0.2412	0.3071
MMVAE ($K = 1$)	0.0354	0.1717	0.2362
Hölder ($K = 1$)	0.0457	0.3252	0.4007
MMVAE+ ($K = 1$)	0.0501	0.1940	0.2884
CMVAE ($K = 1$)	0.0827	0.2456	0.2803
Hölder+ ($K = 1$)	0.0597	0.3510	0.4789
Hölder++ ($K = 1$)	0.0617	0.3518	0.4759

Experimental Results

Improve image quality in generation

We **DiffuseVAE** as a **post-hoc refinement** process. This **improves image quality** without modifying the underlying characteristics of the model.



(a) Hölder++ output

(b) + DiffuseVAE output

Conclusion

UNIVERSITÄT DES SAARLANDES

Our contributions are **threefold**:

(i) **Hölder:** The first exact symmetric Hölder pooling as aggregation.

(ii) **Hölder+:** Learning shared and private subspaces.

(iii) **Hölder++:** Disentangling private and shared subspaces through hierarchical inference.

Experiments show that Hölder++ consistently

- **improves the quality-coherence trade-off,**
- learns more **structured latent spaces**, and
- infers shared representations useful for **downstream tasks**.

Hölder++: Improving the Quality-Coherence Trade-off in Multimodal VAEs

Thank you!

Huyen Vo



María Martínez-García



Isabel Valera



European Research Council
Established by the European Commission

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

UNIVERSITÄT
DES
SAARLANDES

ICML
International Conference
On Machine Learning