MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

European Research Council
Established by the European Commission

UNIVERSITÄT
DES
SAARLANDES

ICML
International Conference
On Machine Learning

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

Presenter: Huyen Vo

Huyen Vo

María Martínez-García

Isabel Valera

Motivation

UNIVERSITÄT DES SAARLANDES

Cross-generation from modality 3 to modality 2 on PolyMNIST

wrong

✅ high quality, ❌ low coherence

blurred

❌ low quality, ✅ high coherence

Our Goal

✅ high quality, ✅ high coherence

Our contributions

UNIVERSITÄT DES SAARLANDES

1. **Hölder:** The first exact implementation of symmetric Hölder pooling as aggregation

2. **Hölder+:** Learning shared and private subspaces

3. **Hölder++:** Disentangling private and shared subspaces through hierarchical inference

Related works in Multimodal VAEs

- **Multimodal VAEs** aggregate unimodal encoders to infer shared latent variables
- [1] views multimodal aggregation as probabilistic opinion pooling

  → Existing works are special cases of **Hölder pooling**

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

$\alpha \to 0$	$\alpha = 0.5$	$\alpha = 1$
Product of experts (**PoE**)	**Symmetric** Hölder pooling	Mixture of experts (**MoE**)
✅ high quality, ❌ low coherence	✅ high quality ✅ high coherence	❌ low quality, ✅ high coherence

[1] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger multimodal Variational Autoencoders.

Hölder: The exact (symmetric) Hölder pooling as aggregation

UNIVERSITÄT DES SAARLANDES

- [1] **HELVAE**: approximates the symmetric Hölder pooling via moment matching
- We instead provide **the first exact implementation** of symmetric Hölder pooling

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

→ **HELVAE** and **Hölder** suffer from limited image diversity

→ **Private latents** capture variations that are not encoded in the shared latent space

[1] AISTATS 2026. Huyen Vo, Isabel Valera. Hellinger multimodal Variational Autoencoders.

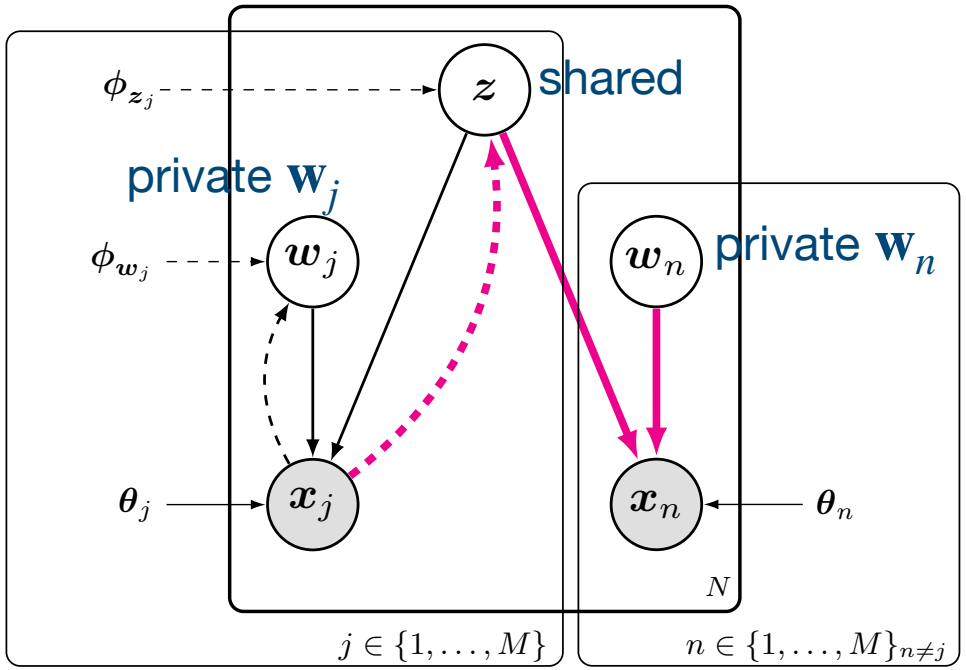
Hölder+: Learning shared and private subspaces

We **factorize** the latent space into shared latent $\mathbf{z}$ and private latents $\mathbf{W}$ subspaces

Shortcut prevention [2]: when $\mathbf{z} \sim q_{\phi_j}(\mathbf{z} \mid \mathbf{x}_j)$,

each $\mathbf{w}_n$ is samples as

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



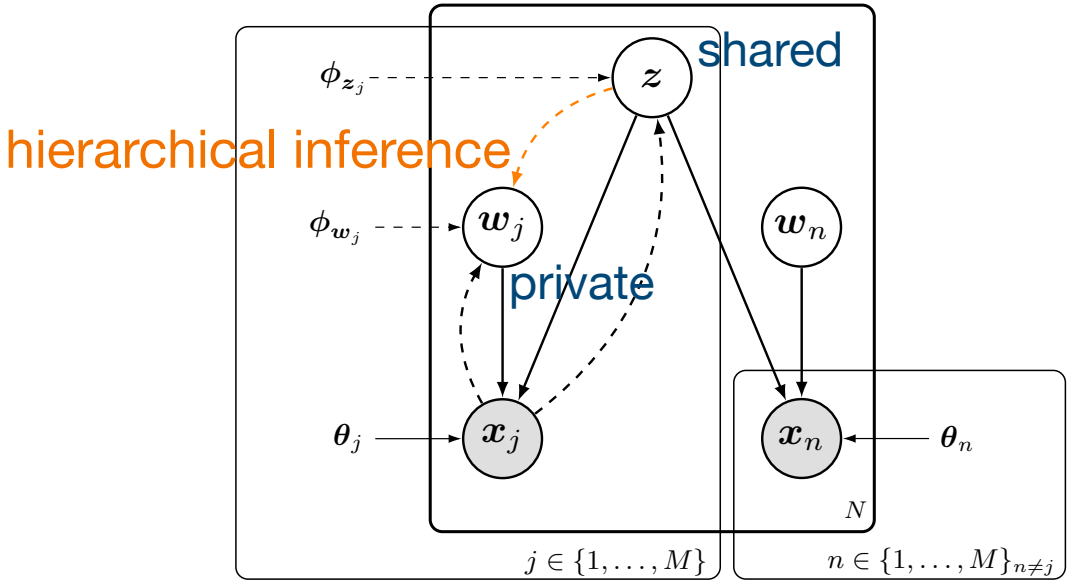
[2] ICLR 2023. Emanuele et al. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises.

Hölder++: Disentangling private-shared subspaces

We encourage disentanglement between the shared and private latents **by design**

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

hierarchical inference



Unimodal components in the Hölder++ objective

$\{\mathbf{w}_j\}_{j=1}^M$ models private (residual) information not captured by $\mathbf{z}$

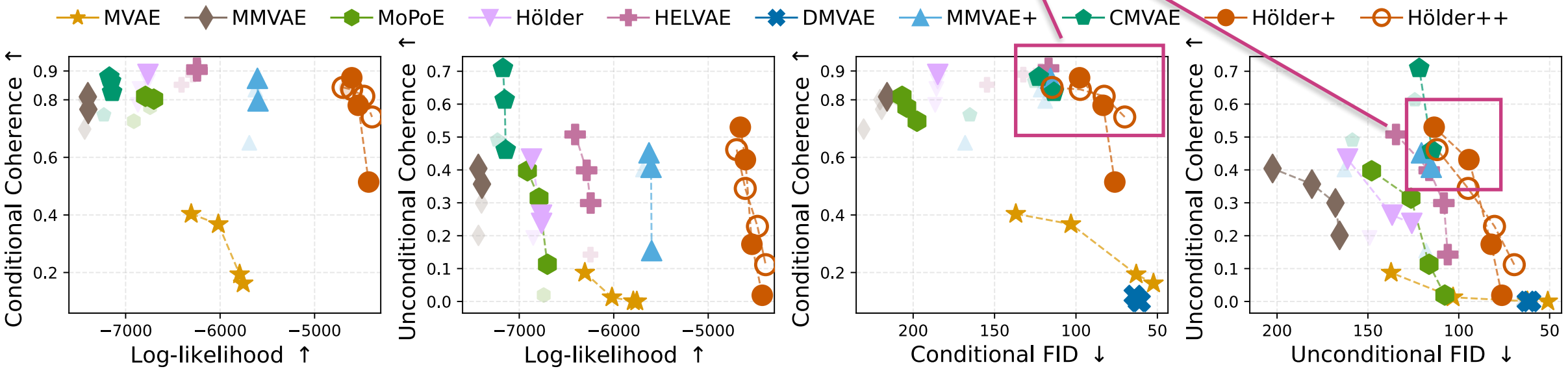
→ **Increase disentanglement** between $\mathbf{z}$ and $\{\mathbf{w}_j\}_{j=1}^M$

Experimental Results: Generative quality and coherence trade-off

UNIVERSITÄT DES SAARLANDES

Hölder+/++ lies close to the Pareto frontier,

outperforming the SOTA **MMVAE+**



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

Experimental Results: Disentanglement of shared and private subspaces

UNIVERSITÄT DES SAARLANDES

Three new disentanglement metrics

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

Method	w content accuracy ↓	z content stability ↑	z content accuracy ↑
MMVAE+	0.138 ± 0.013	0.846 ± 0.166	0.624 ± 0.049
DCMEM	0.321 ± 0.081	0.207 ± 0.110	0.204 ± 0.123
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

Hölder++ (hierarchical inference) **improves disentanglement** between shared and private latents over **Hölder+**

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

European Research Council
Established by the European Commission

ICML
International Conference
On Machine Learning

UNIVERSITÄT
DES
SAARLANDES

Poster Session 6, Thu 9 Jul 10:30 - 12:15

Huyen Vo

María Martínez-García

Isabel Valera