```
Dorent R, Haouchine N, Golby A, Frisken S, Kapur T, Wells W. Unified
Cross-Modal Medical Image Synthesis with Hierarchical Mixture of
Product-of-Experts. IEEE TPAMI, 2025. arXiv:2410.19378
```

Contents

1 One-Sentence Verdict

2 Research Question & Background Gap

- 2.1 Problem
- 2.2 Background Gap

3 Methods & Data

- 3.1 MMHVAE Architecture
- 3.2 Network Architecture
- 3.3 Data & Experimental Design
- 3.4 Segmentation Pipeline Details

4 Key Evidence

- 4.1 Synthesis Quality (Table 1, Fig.3)
- 4.2 Segmentation Downstream (Table 4 — Most Critical)
- 4.3 Registration Downstream (Table 3)
- 4.4 Ablation Studies (Table 2)

5 Author Claims & My Critical Assessment

- 5.1 What the Paper Explicitly States
- 5.2 What Can Be Reasonably Inferred
- 5.3 What Remains Uncertain

6 Relevance to My Project

7 My Questions & Ideas

8 Key References

1 One-Sentence Verdict

The journal upgrade of MHVAE — supports 4 modalities, handles incomplete training data, and for the first time directly validates synthesis quality on downstream tasks (segmentation + registration). **An nnU-Net trained on synthetic data achieves 73.6% Dice on**

iUS brain tumor segmentation, matching the fully-supervised baseline. Deep read, core engine upgrade for Route B.

2 Research Question & Background Gap

1. 2.1 Problem

How to build a unified cross-modal medical image synthesis framework that handles **arbitrary missing modality combinations** (iUS / T2 / ceT1 / FLAIR) during both training and inference, generating sufficiently high-quality synthetic images to support downstream clinical tasks?

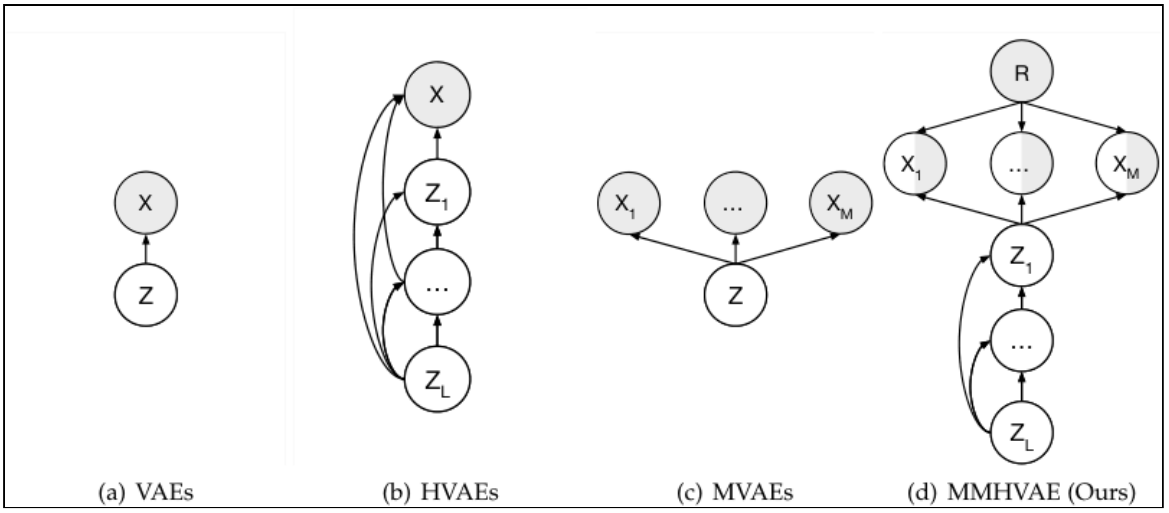
2. 2.2 Background Gap

- Existing unified synthesis methods (MM-GAN, ResViT) handle missing data with zero-filling without learning shared latent representations
- Multi-modal VAEs (MoPoE) learn shared representations but produce blurry images due to low-dimensional latent spaces
- Diffusion models (M2DN) are computationally expensive (18-step inference) and require large datasets to prevent memorization
- **MHVAE (conference version) only handles 2 modalities and requires complete training data** — not all ReMIND patients have complete MRI sequences, creating a practical bottleneck
- All existing methods **evaluate synthesis quality using pixel-level metrics only**, without validating actual value for downstream tasks

3 Methods & Data

3. 3.1 MMHVAE Architecture

MMHVAE = **Mixture of Multimodal Hierarchical VAE**, combining advantages of three VAE types (§2, Fig.1):



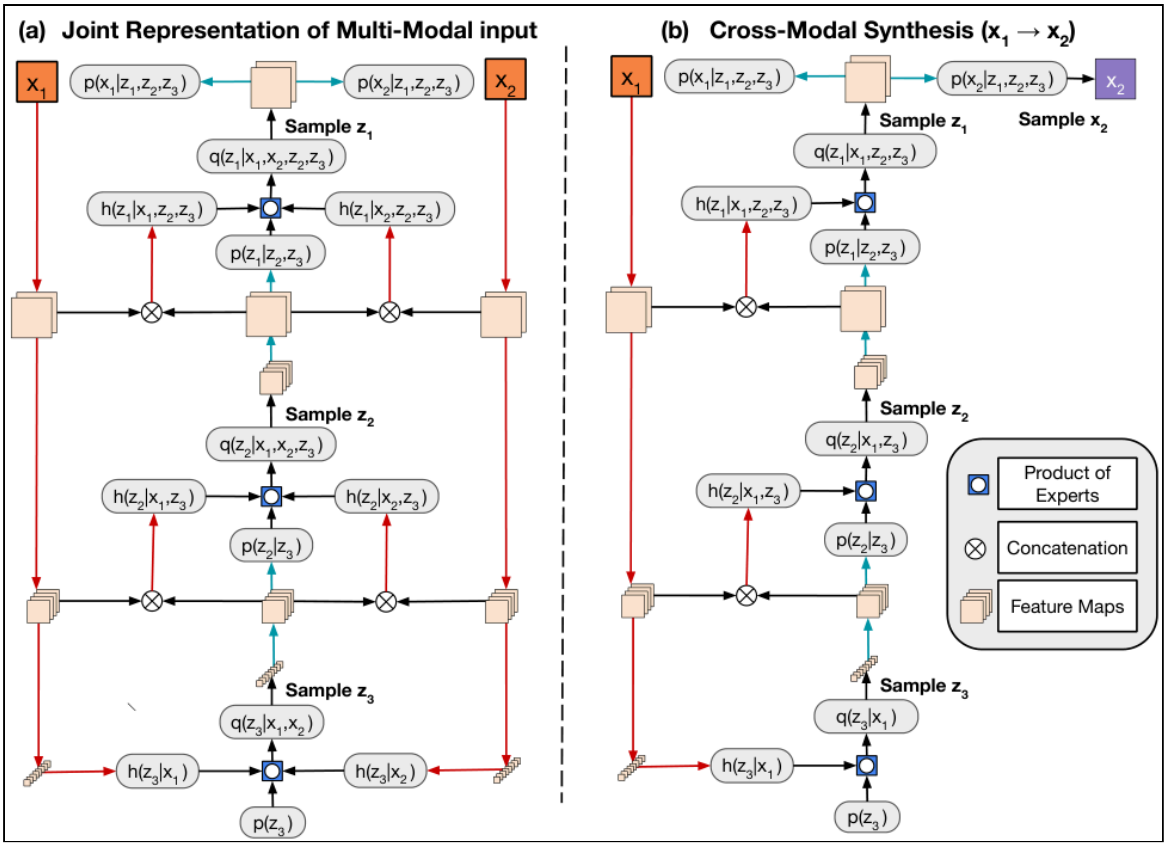
MMHVAE graphical model comparison

MMHVAE's graphical model (d) has both hierarchical structure (from HVAE) and multi-modal conditional independence (from MVAE), introducing a missingness indicator variable R (semi-gray node).

Component	MHVAE (conference)	MMHVAE (this paper)
Modalities	2 (MRI + iUS)	4 (iUS + T2 + ceT1 + FLAIR)
Hierarchy depth	L=7	L=7
Posterior form	Product-of-Experts	**Mixture of Product-of-Experts**
Training data requirement	Complete pairs	**Allows incomplete**
GAN regularization	None	Yes (adversarial constraint on missing modalities)
Parameters	~14M	~14M (only 4% increase)
Inference time	~55ms	~55ms

The mathematical intuition behind the core improvement (§3.3): each component of the MoPoE posterior $q(\mathbf{z}|\mathbf{x}_{\mathbf{r}'}^o)$ is proven to equal the full posterior $p(\mathbf{z}|\mathbf{x}_{\mathbf{r}}^o)$ at ELBO optimality. This means **even with a single input modality, the encoder is encouraged to simultaneously encode observed information and estimate missing information** — the key to cross-modal synthesis (§3.3, Appendix 1.2).

4. 3.2 Network Architecture



MMHVAE network architecture

Left (a): When jointly encoding all modalities, each level fuses features via Product-of-Experts. Right (b): During cross-modal synthesis ($x_1 \rightarrow x_2$), only x_1's encoder is active, but hierarchical sampling still generates x_2.

5. 3.3 Data & Experimental Design

Dimension	Content
Synthesis task data	ReMIND dataset, 3-fold cross-validation
Segmentation downstream	Training: UPenn-GBM (N=611) MRI → virtual sweeps → MMHVAE synthetic iUS + MRI annotations
	Testing: RESECT-SEG (N=22) + ReMIND subset (N=6)
Segmentation model	5-fold nnU-Net ensemble
Registration downstream	Simulated rigid deformations (0–16mm), comparing direct cross-modal vs synthesis-assisted registration
Competing methods	MoPoE, MM-GAN, ResViT, M2DN (diffusion), MHVAE
Evaluation metrics	Synthesis: PSNR, SSIM, LPIPS; Segmentation: DSC, ASSD; Registration: TRE

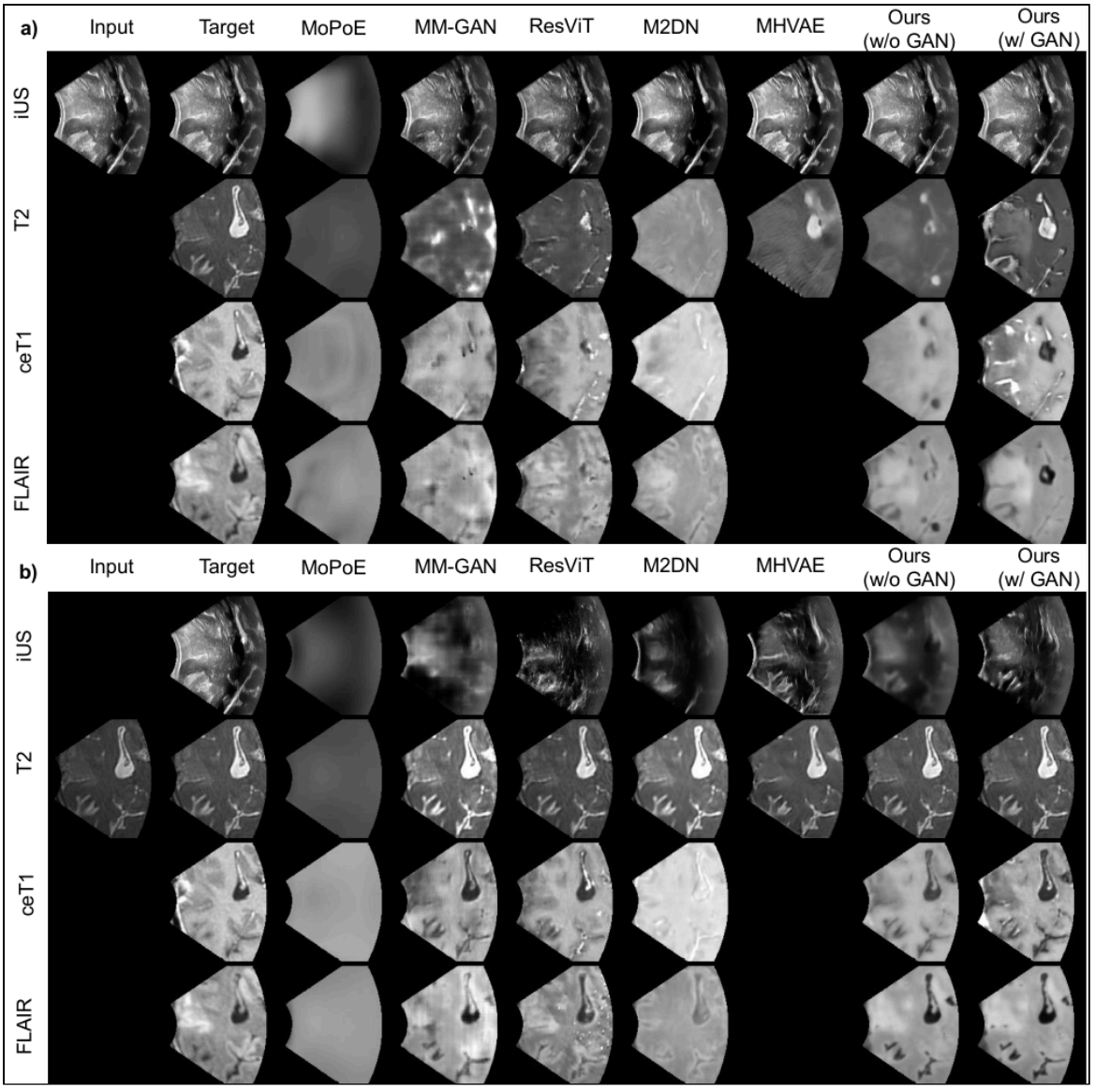
6. 3.4 Segmentation Pipeline Details

The downstream segmentation pipeline is identical to Dorent 2024 (MICCAI):

1. Generate virtual 3D ultrasound sweep paths from UPenn-GBM MRI
2. Synthesize 2D iUS slices along paths using MMHVAE
3. Automatically transfer MRI tumor annotations to synthetic iUS space
4. Train nnU-Net with synthetic iUS + annotations
5. Evaluate on real iUS (RESECT-SEG / ReMIND)

4 Key Evidence

7. 4.1 Synthesis Quality (Table 1, Fig.3)



MMHVAE synthesis quality comparison

Qualitative comparison: synthesizing all MRI modalities from iUS (top) and all modalities

from T2 (bottom). MoPoE is severely blurry; ResViT and M2DN have reasonable structure but distorted textures; MMHVAE (w/ GAN) best preserves speckle texture and anatomical structure.
Quantitatively, MMHVAE achieves SSIM > 98% on **reconstruction/harmonization tasks** (near-perfect) and leads most metrics on **cross-modal synthesis**. Note that MRI→iUS cross-modal synthesis remains the hardest direction (PSNR ~20–21 dB), but MMHVAE has the best LPIPS consistency.

8. 4.2 Segmentation Downstream (Table 4 — Most Critical)

Method	Training Data	RESECT-SEG Dice	ReMIND Dice
MM-GAN synthetic	UPenn MRI→synthetic iUS	53.3%	57.9%
ResViT synthetic	UPenn MRI→synthetic iUS	67.3%	74.7%
M2DN synthetic	UPenn MRI→synthetic iUS	65.2%	62.2%
MMHVAE synthetic	UPenn MRI→synthetic iUS	**73.6%**	**77.6%**
Fully supervised	RESECT-SEG real iUS	—	73.4%
Expert annotation agreement	—	—	84.2%

Warning The fully-supervised baseline comparison requires careful interpretation: it trains on RESECT-SEG (N=22 real iUS) and tests on ReMIND (N=6); while synthetic models train on UPenn-GBM (N=611 MRI→synthetic iUS) and test on RESECT-SEG and ReMIND. Training set sizes and distributions differ, so this is not a strictly fair comparison.

Despite this, **a model trained on synthetic data that has never seen real iUS achieves performance comparable to the fully-supervised model** — this is the strongest evidence for Route B's viability.

9. 4.3 Registration Downstream (Table 3)

Using MMHVAE to synthesize iUS into MRI modalities before performing same-modality registration significantly improves cross-modal registration accuracy:

- **ceT1**: TRE from 14.6mm to 3.2mm (8–12mm displacement range, §5.4)
- **T2**: TRE from 2.0mm to 2.0mm (already good, limited improvement)
- **FLAIR**: TRE from 9.4mm to 4.4mm
  This confirms that MMHVAE synthetic images also have practical value in registration tasks, especially for sequences with large contrast differences from iUS like ceT1 and FLAIR.

10. 4.4 Ablation Studies (Table 2)

Three key findings:

- **Hierarchy depth**: L=7 >> L=5 >> L=3 >> L=1 — hierarchical structure is the foundation of high-quality synthesis
- **Fusion method**: MoPoE fusion > averaging > zero-filled concatenation — probabilistic fusion outperforms simple engineering approaches
- **GAN regularization**: Improves LPIPS (perceptual quality) but worsens PSNR/SSIM — traditional metrics are insufficient for evaluating synthesis quality; **downstream task validation is essential**

5 Author Claims & My Critical Assessment

11. 5.1 What the Paper Explicitly States

- MMHVAE achieves best synthesis quality on most metrics on ReMIND
- Synthetic iUS-trained nnU-Net reaches 73.6% Dice on RESECT-SEG and 77.6% on ReMIND
- Synthesis-assisted registration significantly improves TRE for ceT1 and FLAIR
- Computational efficiency far exceeds ResViT (13G vs 274G MACs) and M2DN (55ms vs 290ms)

12. 5.2 What Can Be Reasonably Inferred

Route B (MRI → synthetic iUS → train segmentation model) has been validated as viable. MMHVAE's multi-modal fusion (T2+ceT1+FLAIR→iUS) outperforms single-modality synthesis, consistent with Rasheed 2025's keypoint matching findings. Temperature parameter T=0.5 is a trade-off point between synthesis quality and diversity — needs tuning in practice. GAN regularization's PSNR/SSIM degradation but downstream task improvement shows that intermediate synthetic images looking "good" and being "useful" are not necessarily the same thing.

13. 5.3 What Remains Uncertain

Small evaluation sample sizes: RESECT-SEG N=22, ReMIND N=6, limiting statistical power.

Circular bias not discussed: Training data annotations come from MRI, and evaluation ground truth is also MRI-derived (except RESECT-SEG, which has independent iUS expert annotations), potentially introducing systematic bias.

Per-tumor-type/grade segmentation performance not reported — differences between high-grade vs low-grade glioma, enhancing vs non-enhancing tumors could be substantial.

UPenn-GBM → ReMIND domain shift: Training uses UPenn-GBM MRI (N=611, single-center UPenn, 1.5T/3T mixed) to synthesize iUS; testing uses ReMIND real iUS (BWH center). MMHVAE itself is trained on ReMIND, so synthetic iUS style is biased toward BWH ultrasound equipment. But UPenn's MRI parameters (sequences, resolution, contrast) differ from ReMIND's MRI — how this difference propagates to synthetic iUS quality is not

discussed. Our project uses ReMIND MRI directly for synthesis, eliminating this domain shift, but training data drops from 611 to ~60 cases.

6 Relevance to My Project

Reusable: The complete synthesis-to-segmentation pipeline (virtual sweep → MMHVAE → nnU-Net) directly corresponds to our Route B. The evaluation scheme using RESECT-SEG and ReMIND aligns with our plan. RESECT-SEG 73.6% Dice and ReMIND 77.6% Dice serve as baseline numbers. Temperature T=0.5 is a practical recommendation.

Not directly reusable or needs adaptation: The paper uses UPenn-GBM as the MRI training source; we may use ReMIND's MRI directly (better matched but smaller). Route A+B combination (registration pseudo labels + synthetic data for joint training) was not attempted in this paper — this is a potential innovation point for our project. Circular bias needs explicit handling in experimental design; RESECT-SEG's independent iUS annotations are the only unbiased evaluation.

Consistent settings: ReMIND dataset, nnU-Net, and RESECT-SEG evaluation fully match. The main difference is training MRI source (UPenn-GBM vs ReMIND).

7 My Questions & Ideas

If ReMIND's own MRI (rather than UPenn-GBM) is used to train MMHVAE and synthesize iUS, how would performance change? ReMIND's MRI and iUS are paired (same patient) — could this provide better patient-specific synthesis? Can Route A's registration pseudo labels and Route B's synthetic data serve as joint training signals? For example, first synthesize large volumes of iUS with MMHVAE, then fine-tune with registration pseudo labels.

GAN regularization improves downstream performance but worsens PSNR/SSIM — should downstream segmentation Dice be used as a proxy metric for synthesis quality (instead of traditional PSNR/SSIM)? MMHVAE's temperature parameter controls synthesis diversity; generating multiple synthetic iUS versions at different temperatures could serve as a data augmentation strategy.

8 Key References

- **[29] MHVAE (MICCAI 2023)**: Conference version, already deep-read. MMHVAE is its journal extension
- **[50] Dorent 2024 (MICCAI)**: Patient-specific segmentation framework, already deep-read. Uses MHVAE synthesis; MMHVAE can directly replace it
- **[43] RESECT-SEG**: The only independent ground truth for iUS segmentation evaluation
- **[36] M2DN (diffusion model)**: Current best diffusion baseline, but 5× computational cost with worse performance — indicating diffusion models have no advantage on this task

  #cross-modal-synthesis #segmentation #registration #ultrasound #MRI #T2 #ceT1 #FLAIR #HierarchicalVAE #MixtureOfProductOfExperts #nnUNet #high-priority