

000 UNPAIRED-TO-PAIRED DATA SYNTHESIS: LEARN- 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 UNPAIRED-TO-PAIRED DATA SYNTHESIS: LEARN- ING TO MODEL DISEASE EFFECTS VIA CONTRASTIVE ANALYSIS OF NEUROIMAGING-DERIVED FEATURES

Anonymous authors

Paper under double-blind review

ABSTRACT

Advances in machine learning have enabled the analysis of complex, high-dimensional datasets, yet neuroimaging lags behind due to data privacy and sharing constraints. Synthetic data offers a promising solution for developing and training models. However, synthesizing disease-specific datasets is challenging, as neurological disorders induce progressive changes in the brain that are subtle and often obscured by normal brain variability. Contrastive analysis provides a framework to learn generative factors that deconvolve variation shared between background (e.g., healthy) and target (e.g., diseased) datasets from variation unique to the target, making it particularly effective in capturing as well as modeling subtle disease effects. In this paper, we reformulate this framework for the synthesis of neuroimaging-derived features, specifically brain regional volumes from T1-weighted structural MRI. Given unpaired neuroimaging samples of healthy and diseased participants, we learn to generate paired healthy and disease feature representations that emulate real disease effects. We show that paired synthesis enables fine-grained, individual-level modeling of disease effects, improving downstream analyses, and supporting more precise exploration of disease heterogeneity. We validate the models on both semi-synthetic and real-world brain regional volume datasets, specifically designed to highlight the heterogeneity parsing capability of contrastive analysis. The models are available at: [\[link\]](#).

1 INTRODUCTION

Machine learning and AI systems have facilitated analysis of high-dimensional, complex, multi-modal data enabling rapid development in fields of computer vision and natural language processing. Tools that automate workflows and optimize time and labor have transformed many industries, streamlining processes while advancing productivity (Rashid & Kausik, 2024). Beyond efficiency, these systems also provide deep insights into underlying structures and patterns hidden within high-dimensional data, enabling data-driven decision-making and discovery. While areas such as robotics, conversational AI, and autonomous driving have seen widespread applications of these technologies, progress in healthcare has been more constrained.

Healthcare data, due to the sensitive personal information it contains, is protected under strict data governance and privacy regulations (Condouah et al., 2025). Moreover, acquiring large-scale biomedical datasets such as imaging, fluid biomarkers, and genomics is expensive and requires coordination among multiple stakeholders, as well as rigorous study designs. These complexities not only limit the volume of data that can be collected but also affect its overall quality and consistency. This along with the difficulty of creating annotated datasets has lead to a shift towards techniques that bridge the gap between data scarcity and model development (Patki et al., 2016; Abdalla et al., 2025; Alzubaidi et al., 2023). **Synthetic data** has emerged as a promising approach in this context, providing a means to augment limited datasets, balance class distributions, and generate realistic samples while preserving patient privacy (Rajotte et al., 2022). By capturing the statistical properties of real biomedical data, synthetic datasets can support the development of more robust, generalizable models without direct reliance on sensitive patient records.

054 Various **generative modeling** techniques have been explored for synthetic data generation (Bauer
 055 et al., 2024), including Gaussian Mixture Models (Pezoulas et al., 2022), Kernel Density Estimation
 056 (Chintapalli et al., 2024), Generative Adversarial Networks (GANs) (Che et al., 2017), Variational
 057 Autoencoders (VAEs) (Jeong et al., 2025), among others. These approaches are typically applied
 058 directly to the source dataset, where the goal is to learn its underlying probability density. Once the
 059 density is estimated, synthetic samples can be generated by drawing from the learned distribution.
 060 However, applying these standard approaches to healthcare data is challenging. Healthcare data is
 061 extremely nuanced, exhibiting variability across participants due to factors such as demographics,
 062 genetics, lifestyle, and disease status. Independently modeling patient and control distributions may
 063 fail to preserve the true differences between the groups, since disease effects are often subtle and
 064 intermingled with the natural variability observed in healthy individuals (Marquand et al., 2016).
 065 This challenge is particularly evident in neurological disorders, which are characterized by het-
 066 erogeneous disease effects. Conditions such as Alzheimer’s disease (AD) and schizophrenia show
 067 substantial **heterogeneity** in both clinical presentation and disease progression (Lam et al., 2013;
 068 Picardi et al., 2012). If **paired** observations were available, that is, counterfactual representations
 069 of the same participant with and without disease, studying heterogeneous disease effects would be
 070 straightforward. However, such paired data are inherently unavailable, as participants are either have
 071 the disease or do not. As such, attempts to explain and study this heterogeneity have increasingly
 072 focused on methods that explicitly model how the patient distribution deviates from the healthy con-
 073 trol distribution, while disentangling the disease effects from confounding healthy variability (Dong
 074 et al., 2015; Chand et al., 2020; Yang et al., 2021). These approaches underscore a key principle
 075 for generative modeling of healthcare data: to produce synthetic datasets of healthy and diseased
 076 participants that are meaningful, it is crucial to explicitly separate disease-related variability from
 077 normal inter-individual variability, ensuring that the synthetic data accurately captures the subtle but
 078 systematic differences between the two datasets.

078 Uncovering how two datasets differ is a key focus of **contrastive analysis (CA)**, a subfield of rep-
 079 resentation learning (Louiset et al., 2024). CA aims to learn generative or latent factors that explain
 080 both the variation shared across a background dataset (e.g., healthy participants) and a target dataset
 081 (e.g., diseased participants), as well as the factors that explain variation specific to the target dataset.
 082 Variation shared between the datasets is generally referred to as **shared** or **common** variation, while
 083 variation unique to the target dataset is referred to as **salient** variation. Separating the salient factors
 084 from common factors enables controlled synthesis of background and target samples while also fa-
 085 cilitating the study of variation unique to the target distribution, such as disease-related heterogeneity
 086 (Aglinskas et al., 2022). **Moreover, this framework naturally supports the generation of paired data,**
 087 **making it possible to synthesize a healthy sample alongside its diseased counterpart with per-**
 088 **sonalized disease effects.** Here, the counterfactual healthy sample is a data-driven approximation that
 089 reconstructs an individual using only the shared latent factors while removing salient disease-related
 090 attributes. Importantly, these counterfactuals should not be interpreted as causal in the sense of
 091 structural causal models (Pawlowski et al., 2020); rather, it is a generative, representation-level con-
 092 struct that isolates target-specific variation from common variation. This distinction is important, as
 093 the goal of the CA framework is to disentangle and model differences between datasets, not to infer
 094 causal effects.

094 The CA framework has most commonly been implemented using variational autoencoders (Abid &
 095 Zou, 2019; Severson et al., 2019), where it is commonly referred to as contrastive VAE. In this paper,
 096 we propose to adapt this framework for synthesizing neuroimaging-derived features, specifically
 097 regional brain volumes obtained from T1-weighted structural MRI, of healthy and diseased samples.
 098 We chose regional brain volumes, as they are widely used biomarkers to study changes in brain
 099 structure and atrophy due to neurological disorders, and have been used for diagnosis, prognosis,
 100 and subtype discovery (Eskildsen et al., 2015; Chand et al., 2020). **Although we focus on regional**
 101 **brain volumes, the proposed framework is modality-agnostic: any structured neuroimaging-derived**
 102 **feature vector (e.g., cortical thickness, diffusion-derived connectometry, or functional connectivity)**
 103 **can be encoded into common and salient latent factors.** We restrict our evaluation to anatomical ROI
 104 volumes because (i) they are the most consistently available measurements across large-scale multi-
 105 cohort datasets, (ii) site/scanner harmonization methods for volumes are well established (Pomponio
 106 et al., 2020), and (iii) disease effects are well characterized, enabling clear validation.

106 To better capture variability across individuals, we leverage data aggregated from multiple studies
 107 to build this framework. Our contributions are summarized below:

108 1 **Adaptation of the CA framework for neuroimaging-derived features:** We apply con-
 109 trastive analysis using variational autoencoders for synthesizing regional brain volumes of
 110 both healthy and diseased participants.
 111 2 **Paired synthetic data generation:** The framework supports synthesis of paired healthy-
 112 diseased samples, enabling personalized simulations of disease effects for downstream
 113 analysis. We formally show how CA optimizes an objective analogous to maximizing
 114 the joint likelihood of the healthy and disease distributions.
 115 3 **Isolating disease-related variability in mild cognitive impairment (MCI) and**
 116 **Alzheimer’s Disease (AD):** We show that applying CA to MCI and AD participants cap-
 117 tures disease-related salient variation.
 118 4 **Multi-study integration:** Our approach leverages data aggregated across multiple cohorts,
 119 improving the diversity and utility of the synthesized samples.
 120 5 **Facilitating access to derived features:** By synthesizing neuroimaging-derived features
 121 that are otherwise tedious to obtain due to preprocessing and segmentation requirements,
 122 our method provides a practical and clinically useful alternative data source for research
 123 applications.

125 2 BACKGROUND AND RELATED WORKS

127 This work relates to contrastive analysis, variational inference, synthetic data generation, and parsing
 128 disease heterogeneity.

130 2.1 CONTRASTIVE ANALYSIS

132 Formally, as described in (Abid & Zou, 2019; Severson et al., 2019), given i.i.d. samples $\{x\}_{i=1}^{N_x}$
 133 from a target distribution and samples $\{y\}_{j=1}^{N_y}$ from a background distribution, contrastive analysis
 134 aims to learn latent variables (z, s) , where $z \in R^{D_z}$ captures the **shared variation** between the
 135 target and background datasets, and $s \in R^{D_s}$ captures the **salient variation** that is unique to the
 136 target dataset. Under the variational autoencoder setting, two probabilistic encoders $q_{\phi_z}(z|x)$ and
 137 $q_{\phi_s}(s|x)$ approximate the posteriors over the two sets of latent variables z and s , respectively. Then
 138 it is assumed that each sample x_i or y_j is drawn from a conditional distribution $p_{\theta}(\cdot|z, s)$, parame-
 139 terized by unknown decoder parameters θ . Typically for the background samples, the salient latent
 140 variables are manually set to 0, but modifications exist (Weinberger et al., 2022; Louiset et al., 2023).
 141 To optimize the framework, a variational lower bound is defined for the log-likelihoods of individual
 142 data points, such that:

$$\log p(x_i) \geq \mathbb{E}_{q_{\phi_z}(z), q_{\phi_s}(s)} \log p_{\theta}(x_i|z, s) - KL(q_{\phi_z}(z|x_i) || p(z)) - KL(q_{\phi_s}(s|x_i) || p(s)) \quad (1)$$

$$\log p(y_j) \geq \mathbb{E}_{q_{\phi_z}(z)} \log p_{\theta}(y_j|z, 0) - KL(q_{\phi_z}(z|y_j) || p(z)) \quad (2)$$

145 Here, KL is the Kullback–Leibler divergence and $p(z)$ and $p(s)$ are the prior distributions over the
 146 two sets of latent variables z and s , respectively. With $p(z)$ and $p(s)$ assumed to be multivariate
 147 isotropic Gaussians $\mathcal{N}(0, I)$. The two encoders along with the decoder are trained by maximize the
 148 sum of the objective functions (1) and (2).

150 2.2 PAIRED DATA SYNTHESIS: SHARED OBJECTIVE WITH CONTRASTIVE VAE

152 To synthesize paired samples (x, y) , the joint distribution $p(x, y)$ must first be defined. However,
 153 since in practice, paired data is often unavailable or difficult to acquire, we need to learn the joint
 154 distribution using unpaired observations from the marginal distributions i.e. $x \sim p(x)$ and $y \sim$
 155 $p(y)$. Directly learning the joint distribution $p(x, y)$ from unpaired samples $\{x\}_{i=1}^{N_x}$ and $\{y\}_{j=1}^{N_y}$
 156 is non-trivial since x and y are not independent. But variational inference can impose conditional
 157 independence using latent variables. We again assume that z captures variation common to both $(x,$
 158 $y)$, while s captures target-specific variation. For neuroimaging features, z may encode demographic
 159 or scanner variability shared between healthy and diseased participants, while s encodes disease-
 160 related variability (e.g., atrophy patterns in MCI or AD). Under these assumptions, the conditional
 161 joint distribution can be written as

$$p(x, y|z, s) = p(x|z, s)p(y|z) \quad (3)$$

Similarly, the approximate posterior distribution factorizes as

$$q(z, s|x, y) = q(z|x, y)q(s|x) \quad (4)$$

Using variational inference, the log-likelihood of paired samples can be written as

$$\begin{aligned}
\log p(x, y) &= \mathbb{E}_{q(z, s|x, y)} \log \frac{p(x, y, z, s)}{q(z, s|x, y)} + KL(q(z, s|x, y) || p(z, s|x, y)) \\
&\geq \mathbb{E}_{q(z, s|x, y)} \log \frac{p(x, y|z, s)p(z, s)}{q(z, s|x, y)} \text{ (since KL divergence is non-negative)} \\
&\geq \mathbb{E}_{q(z|x, y)q(s|x)} \log \frac{p(x|z, s)p(y|z)p(z)p(s)}{q(z|x, y)q(s/x)} \text{ (using (3), (4))} \\
&\geq \mathbb{E}_{q(z|x, y)q(s|x)} \log p(x|z, s) + \mathbb{E}_{q(z|x, y)} \log p(y|z) - KL(q(z|x, y) || p(z)) - KL(q(s|x) || p(s))
\end{aligned} \tag{5}$$

This is a variational lower bound for log-likelihood of paired samples. This formulation shows that the joint log-likelihood decomposes into two reconstruction terms (for x and y) regularized by KL penalties on the common and salient latents. Importantly, because the bound depends on $q(z|x, y)$, paired data is still required. However, if we explicitly decouple the posterior such that

$$q(z|x, y) = q(z|x) = q(z|y) \quad (6)$$

then the joint log-likelihood can be maximized using unpaired samples. This essentially means that the common latent can be inferred from either the background or target sample, such that it explains the variation common to both the background and target dataset. **Equation (5) simplifies to:**

$$\begin{aligned} \log p(x, y) &\geq \mathbb{E}_{q(z|x, y)q(s|x)} \log p(x|z, s) + \mathbb{E}_{q(z|x, y)} \log p(y|z) - \frac{1}{2} KL(q(z|x)||p(z)) \\ &\quad - \frac{1}{2} KL(q(z|y)||p(z)) - KL(q(s|x)||p(s)) \end{aligned} \quad (7)$$

We can see that (7) is equivalent to the sum of (1) and (2), which is the objective of CA, albeit with a weaker penalty on the KL divergence terms. This shows that CA can be understood as implicitly maximizing a joint likelihood, even when trained only on unpaired data.

Recent CA methods with variational autoencoders already enforce this decoupling to some capacity. In (Weinberger et al., 2022), a maximum mean discrepancy (MMD) term is added to the overall objective so that the distribution of the common latent is same across target and background samples. Similarly in (Luiset et al., 2024), a mutual information term is added to the objective to maximize the mutual information between the common latent and the two datasets.

3 METHODS: CONTRASTIVE ANALYSIS FRAMEWORK FOR UNPAIRED-TO-PAIRED DATA SYNTHESIS

Building on the contrastive VAE framework detailed in Section 2.1, we develop a model for generating paired neuroimaging-derived features from unpaired healthy and disease samples. Our approach encodes each sample into a common latent z , capturing variation shared across datasets, and a salient latent s , capturing disease-specific variation. By explicitly decoupling these latent spaces, the model can leverage unpaired data while enabling the generation of realistic paired samples.

3.1 MODEL ARCHITECTURE

The general architecture of the model is illustrated in Figure 1. The model consists of two probabilistic encoders, q_{ϕ_z} and q_{ϕ_s} , mapping input features into the common and salient latent spaces, respectively, and a shared decoder p_{θ} that reconstructs input features from the concatenated latent representations. For background (healthy) samples, the salient latent is fixed to zero, whereas both latents are inferred for target (disease) samples. The latent dimensionality and encoder/decoder architectures are chosen to balance reconstruction fidelity with latent disentanglement. More on this in the implementation details Section 4.2

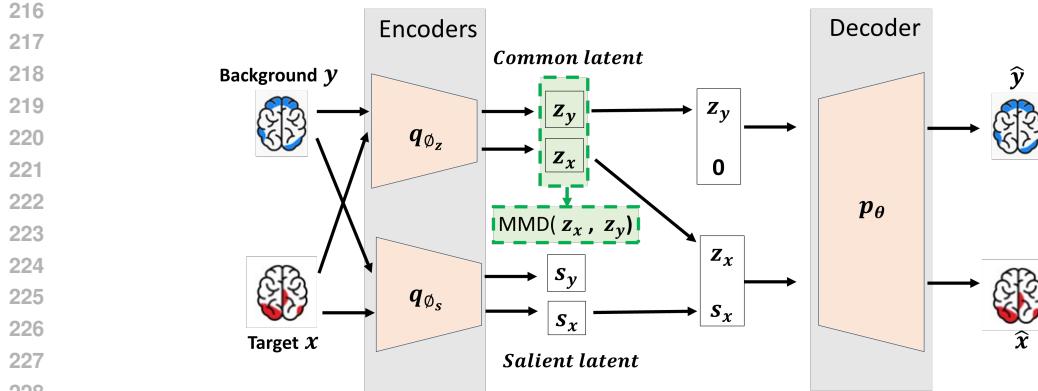


Figure 1: Illustration of the contrastive VAE model. The model separates latent variables into (i) common latents z , which capture variation shared between background (e.g., healthy) and target (e.g., diseased) data, and (ii) salient latents s , which capture variation unique to the target. Two encoders, q_{ϕ_z} and q_{ϕ_s} , map inputs into these respective representations, which are concatenated and passed to a decoder p_{θ} for reconstruction. For background samples, the salient latents are fixed to zero. To enforce decoupling of the posterior distribution of the common latent, such that this latent captures shared variation across the target and background samples, maximum mean discrepancy between common latent representations of target (z_x) and background (z_y) is minimized during model training. Synthetic data are generated by sampling $z, s \sim \mathcal{N}(0, I)$ and decoding: paired target samples use both z and s , paired background samples use z with $s = 0$, and unpaired background samples use z' sampled independently from z .

3.2 TRAINING OBJECTIVE

The model is trained to maximize a variational lower bound on the log-likelihood of target and background samples (see equation (1), (2)). The training objective combines reconstruction terms, KL penalties on the latent posteriors, and a regularization term to align common latent distributions across datasets. This regularization term penalizes the MMD between z_x and z_y which are the common latent representation of the target and background samples, respectively. By doing so, we ensure that the common latent represents the variations shared between the target and background samples, ensuring that the common posterior under the paired setting is decoupled (see equation 6). So, the overall contrastive VAE objective that needs to be maximized is

$$\mathcal{L} = \sum_{i=1}^{N_x} \mathcal{L}_x(x_i) + \sum_{j=1}^{N_y} \mathcal{L}_y(y_j) - \lambda \text{MMD}(z_x, z_y) \quad (8)$$

where $\mathcal{L}_x(x_i) = \mathbb{E}_{q_{\phi_z}(z)q_{\phi_s}(s)} \log p_{\theta}(x_i|z, s) - KL(q_{\phi_z}(z|x_i)||p(z)) - KL(q_{\phi_s}(s|x_i)||p(s))$ is the variational lower bound for log-likelihood of target samples (see equation (1)), $\mathcal{L}_y(y_i) = \mathbb{E}_{q_{\phi_z}(z)} \log p_{\theta}(y_i|z, 0) - KL(q_{\phi_z}(z|y_i)||p(z))$ is the variational lower bound for log-likelihood of background samples (see equation (2)), and λ controls the strength of the alignment penalty. This objective encourages separation of common and salient latents while ensuring that the common latent captures variation shared across healthy and diseased participants.

3.3 SYNTHETIC PAIRED DATA GENERATION

Once the model is trained and the parameters ϕ_z , ϕ_s , and θ are learned, synthetic samples can be generated by sampling from the latent priors.

- 1) **Synthesizing paired background and target samples using latent priors:** Target samples are synthesized by decoding (z, s) , with $z \sim \mathcal{N}(0, I)$ and $s \sim \mathcal{N}(0, I)$. Paired background samples are synthesized by decoding $(z, s = 0)$.

270
 271 2) **Conditional synthesis of paired background and target samples:** If unpaired real target
 272 and background samples are available, the paired counterpart can be synthesized condition-
 273 ally. The common latent z is inferred from the given sample, and to synthesize a paired
 274 target sample, $s \sim \mathcal{N}(0, I)$ is sampled and decoded with z : $p(x | z, s)$, and to synthesize a
 275 paired background sample, s is set to 0 and decoded with z .

276 **4 EXPERIMENTS**
 277

278 To implement the framework for synthesizing paired neuroimaging derived features from unpaired
 279 neuroimaging derived features, we first applied the framework to semi-synthetic data to validate its
 280 performance under controlled environment and then applied it to real data.
 281

282 **4.1 DATASETS**
 283

284 Both the semi-synthetic and real-data experiments used data from the iSTAGING consortium (Habes
 285 et al., 2021), which consolidated and harmonized imaging and clinical data from multiple studies
 286 spanning a wide age range (22–90 years). The consortium includes neuroimaging, demographic,
 287 and clinical measures from participants clinically diagnosed as cognitively normal (CN), with mild
 288 cognitive impairment (MCI), or with Alzheimer’s disease (AD). The neuroimaging-derived features
 289 consist of 139 anatomical brain region-of-interest (ROI) volumes (119 gray matter ROIs and 20
 290 white matter ROIs) extracted from baseline T1-weighted MRI scans using a multi-atlas label fusion
 291 method (Doshi et al., 2016). To harmonize data across studies and mitigate site effects, ComBat-
 292 GAM harmonization (Pomponio et al., 2020) was applied to the ROI volumes while adjusting for
 293 age, sex, and intracranial volume (ICV).

294 **Semi-Synthetic Experiment:** For this experiment, 3,900 cognitively normal participants (Age 62.9 ± 7.6 years, 52% female) were sampled from the iSTAGING dataset. Of these, 1,950 samples were
 295 used to construct a synthetic disease cohort with heterogeneous disease effects. Heterogeneity was
 296 simulated by defining three subtypes, each with 10–30% atrophy in predefined ROIs, with some
 297 overlap between subtypes (see Table 1).**We applied 10–30% synthetic atrophy in the semi-synthetic**
 298 **data to reflect realistic disease effects observed in neurodegenerative conditions, ensuring clinically**
 299 **plausible variation while providing a meaningful challenge for generative modeling**(Whitwell,
 300 2010). The procedure was as follows: first, the 139 ROI volumes were covariate-corrected for age,
 301 sex, and ICV (volumes were residualized using linear regression). Then, for each synthetic disease
 302 sample k , one of the three subtypes was randomly assigned. For each ROI l belonging to the chosen
 303 subtype, the corresponding volume v_{kl} was reduced according to

$$v_{kl} = v_{kl} - \text{Uniform}(0.1, 0.3) \times v_{kl}.$$

306 Finally, the covariate effects were added back to preserve population-level structure. The cogni-
 307 tively normal and synthetic disease participants were respectively split into train (N=1050), cross-
 308 validation (N=450), and test (N=450) cohort for training and validating the contrastive VAE model.
 309 All volumes were standardized with respect to the cognitively normal training set.

310 **Real Data Experiment:** For the real-data experiment, we sampled 5,212 cognitively normal partic-
 311 ipants (Age 64.8 ± 10.5 years, 56.8% female) and 1,409 participants clinically diagnosed with MCI
 312 or AD (Age 77.0 ± 9.1 years, 53.7% female) from the iSTAGING dataset (independent from the
 313 semi-synthetic experiment). The healthy and diseased dataset were respectively split into training,
 314 cross-validation, and test cohorts using a 70:15:15 ratio. Because the healthy and disease cohorts
 315 have different demographic distributions, ROI volumes were covariate-corrected for age, sex, and
 316 ICV **and the residuals were used in training the model**. Specifically, volumes were residualized using
 317 linear regression parameters estimated from the cognitively normal training cohort. This procedure
 318 prevents the model from conflating demographic differences with disease-related effects. All ROI
 319 volumes were standardized with respect to the cognitively normal training set.

320 To further assess whether the trained model generalizes and captures disease-related salient varia-
 321 tion, we evaluated it on an out-of-distribution (OOD) dataset from a study that was held-out from
 322 training. For this, we used the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset (Toga
 323 & Crawford, 2015) which has comprehensive clinical and cognitive measures that quantify disease
 burden. We selected 2,437 participants across ADNI1/2/GO/3 with baseline regional brain volumes.

324
 325 Table 1: Bilateral ROIs included in disease patterns (OP:opercular part) in Semi-synthetic experi-
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377

ment, 10-30% atrophy simulated

ROI	Atrophy Patterns		
	S1	S2	S3
Amygdala	✓		
Temporal pole	✓	✓	✓
Hippocampus	✓	✓	
Angular gyrus			✓
Entorhinal area	✓		
Frontal operculum		✓	
Inferior temporal gyrus		✓	
Lateral orbital gyrus			✓
Medial frontal cortex	✓	✓	
Middle frontal gyrus	✓		
Middle occipital gyrus			✓
OP of the inferior frontal gyrus			✓
Parahippocampal gyrus		✓	
Posterior insula		✓	
Parietal operculum	✓		✓
Supramarginal gyrus			✓

This dataset comprised of 871 cognitively normal (Age 72.6 ± 6.3 years, 41.8 % female), 1,101 MCI (Age 72.8 ± 7.6 years, 56.6% female), and 420 AD (Age 74.9 ± 7.8 years, 44% female) samples.

4.2 IMPLEMENTATION DETAILS

Since the inputs are one-dimensional regional brain volume features, both the common and salient encoders as well as the decoder were implemented as shallow networks consisting of two linear layers with LeakyReLU activations. Empirically we found that a latent dimensionality of 5 for both the common and salient variables provided sufficient capacity to capture the variability across the target and background datasets. The regularization coefficient λ was set to 10 for the semi-synthetic experiments and to 100 for the real-data experiments. Max epochs is 500. For optimization, we used ADAM optimizer with learning rate 0.001. β_1 and β_2 are 0.5 and 0.999, respectively. To monitor the quality of synthesized healthy and diseased data during training, in addition to the reconstruction loss, we monitor the Wasserstein distance between the synthesized and cross-validation data distributions. Specifically, we use the 2-Wasserstein distance formula for two multivariate gaussian measures(Mallasto et al., 2022) to assess whether the generated data distribution is close to the observed ground truth reference distribution. We trained the model until the maximum number of epochs, unless a worsening trend was observed in the Wasserstein distance. **Because the decoder outputs continuous ROI volumes, we model $p_\theta(\cdot | z, s)$ as a multivariate Gaussian distribution with fixed identity covariance. Under this assumption, the decoder predicts only the mean of the distribution, and the reconstruction terms in the objective reduce to the mean squared error (MSE) between the input and the reconstructed output.** This is a standard choice for continuous features, and it enables the model to focus on capturing the mean structure of the data while ensuring stable and well-behaved training.

5 RESULTS

We first evaluated the contrastive VAE model on the semi-synthetic dataset to assess its ability to capture disease-specific effects. As shown in Figure 2a, the model generates paired data that reflect the simulated disease patterns. Specifically, bilateral hippocampal atrophy was introduced in subtypes S1 and S2, and the generated paired samples successfully replicate these effects. The latent representations further support these findings. Figure 2b shows that the common latent dimensions are well-aligned between healthy and disease samples, while the salient latent dimensions encode subtype-specific information. In particular, the first four salient dimensions provide a structured representation of subtypes S1–S3.

To assess the contribution of the contrastive structure and the MMD alignment term, we performed an ablation study (see Appendix A.1 for full details). Table A1 shows that the CA-VAE with MMD regularization generates synthetic target and background samples with higher fidelity compared to

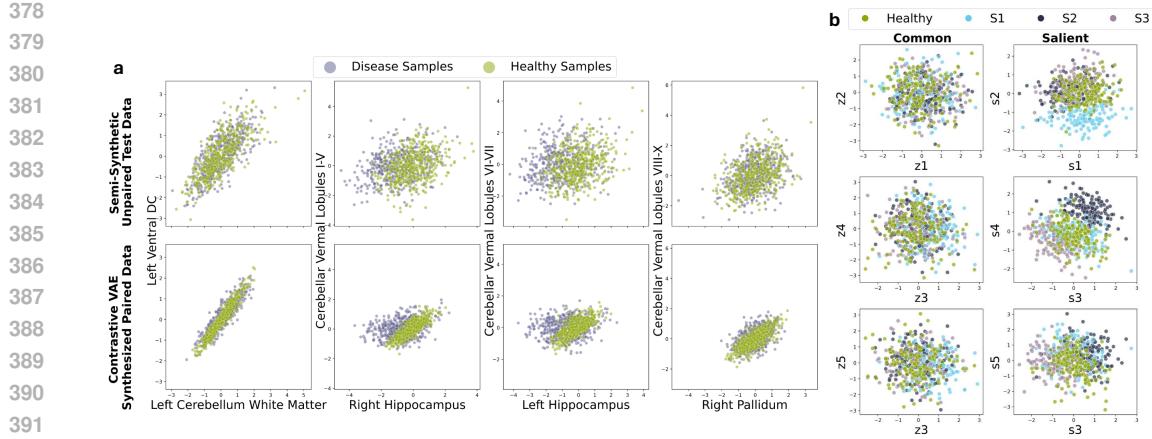


Figure 2: Semi-Synthetic experiment results: (a) Distribution of standardized brain volumes in the semi-synthetic test dataset (top row) and the contrastive VAE-generated paired dataset (bottom row). The bilateral hippocampus is included in subtypes S1 and S2. (b) Scatterplot of common and salient latent means when the semi-synthetic test dataset is input to the encoders. $\{z_1, \dots, z_5\}$ and $\{s_1, \dots, s_5\}$ denote the five common and salient dimensions, respectively.

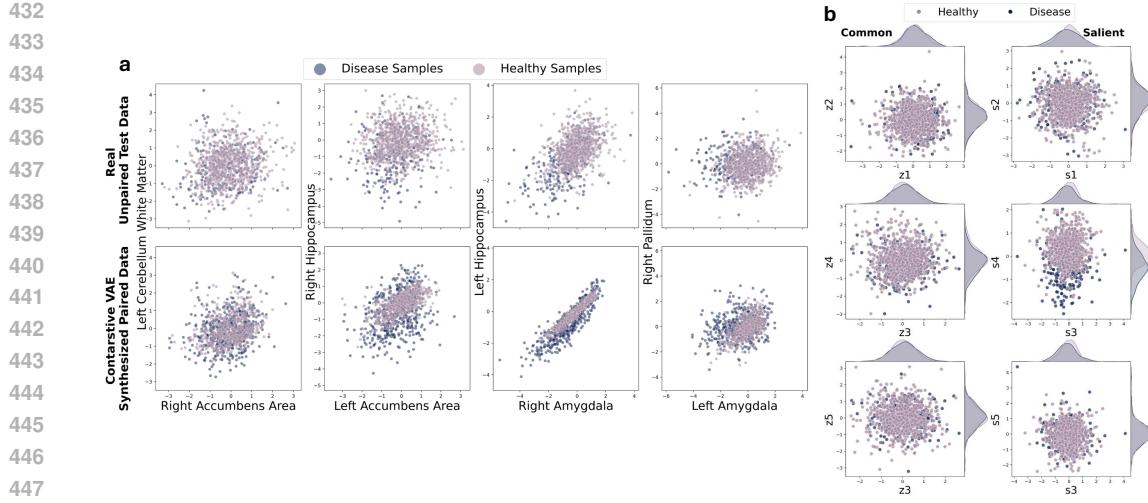
a standard conditional VAE and to variants without MMD. Increasing the MMD penalty improves latent-space decoupling, as measured by a logistic regression classifier on the common latent means (AUC closer to 0.5 indicates better decoupling, Figure A1a). Additionally, K-means clustering on the salient latents demonstrates that the CA-VAE with $MMD \lambda = 10$ captures disease-subtype heterogeneity more effectively than the baseline model (Figure A1b). These results confirm that both the contrastive structure and the MMD alignment term are important for generating decoupled and subtype-informative latent representations.

After validating the model in a controlled setting, we trained it on real data. Figure 3a shows that real data are noisy, and even after covariate correction, the distributions of healthy and disease samples exhibit substantial overlap. Despite this, the model learns to synthesize paired healthy and disease samples that highlight structural variations in the bilateral hippocampus and amygdala, regions well-documented in the AD literature (Qu et al., 2023). Additionally, the common latent representation remains well-aligned across samples, while salient dimension 4 captures disease-related variation (Figure 3b). In this clinical cohort, the common latent dimensions were generally well-aligned across healthy and disease samples; however, common dimension 4 showed some separation between CN and AD subjects, suggesting subtle differences captured in the shared latent space. Given that ADNI is a clinical cohort, some confounding differences between healthy and disease participants are possible. As expected, salient dimension 4 captured variation associated with MCI and AD. Overall, the salient latent space captures disease severity: CN and MCI samples show considerable overlap, whereas AD samples are more distinct (Figure 4a), consistent with MCI representing early cognitive symptoms that may progress to AD.

We further leveraged the model’s conditional synthesis capability to generate counterfactual healthy samples for each individual in the ADNI dataset. By computing the mean absolute difference (MAD) between observed and synthesized volumes, we observed a clear trend corresponding to disease severity: MAD was lowest for CN, higher for MCI, and highest for AD (Figure 4b). This demonstrates that the model has learned to disentangle shared from disease-related variation and can provide quantifiable measures of disease severity at the individual level.

6 DISSCUSSION

In this paper, we adapted a contrastive VAE framework for generating synthetic neuroimaging-derived brain regional volumes, providing a generative model that can help overcome challenges of limited sample sizes, imbalanced datasets, and privacy constraints. By making this model available,



448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

Figure 3: Real data experiment results: (a) Distribution of standardized brain volumes in the real test dataset (top row) and the contrastive VAE-generated paired dataset (bottom row). Atrophy can be seen in bilateral hippocampus and amygdala. (b) Scatterplot of common and salient latent means when the real test dataset is input to the encoders. $\{z_1, \dots, z_5\}$ and $\{s_1, \dots, s_5\}$ denote the five common and salient dimensions, respectively.

we aim to support research, education, and broader access to neuroimaging data, which often has a steep learning curve and significant preprocessing demands.

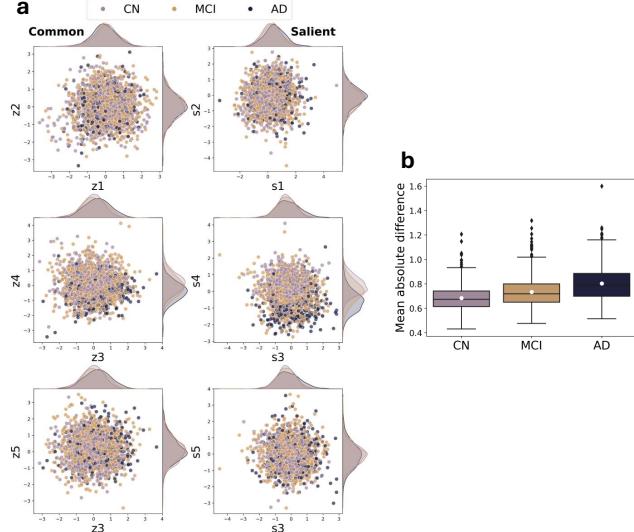


Figure 4: Results from OOD ADNI dataset: (a) Scatterplot of common and salient latent means when the ADNI dataset is input to the encoders. $\{z_1, \dots, z_5\}$ and $\{s_1, \dots, s_5\}$ denote the five common and salient dimensions, respectively. b) Boxplot of the mean absolute difference between observed brain volumes and contrastive VAE-synthesized healthy brain volumes. Healthy volumes were generated by combining each sample's inferred common latent representation with a salient latent set to zero, which is then passed through the decoder.

We showed that, under certain conditions, the objectives of contrastive analysis and paired data synthesis are equivalent, and used maximum mean discrepancy to achieve this alignment. By generating paired healthy and diseased samples, the model enables personalized modeling of disease effects, allowing fine-grained exploration of individual differences. This approach can be leveraged

486 for studying disease heterogeneity and subtype discovery. We demonstrated this capability on OOD
 487 ADNI data where salient latent dimensions captured variation associated with MCI and AD, re-
 488 reflecting disease-specific effects, while common dimensions preserved shared anatomical structure.
 489 Moreover, conditional synthesis of counterfactual healthy samples revealed an increasing trend in
 490 mean absolute differences between observed and counterfactual volumes across CN, MCI, and AD
 491 participants. This shows that the generalized well to OOD data and can be used to quantify disease
 492 severity at an individual-level.

493 Limitations of our current approach include its dependence on input data quality and the exclusive
 494 focus on regional brain volumes, which may not capture all aspects of disease effects. The model
 495 was evaluated primarily on a single dataset, and applying it to multiple cohorts could further test
 496 its generalizability. Additionally, incorporating regularization strategies based on known disease
 497 effects could further improve the fidelity of synthetic data. While the current latent representations
 498 are interpretable, they do not explicitly model discrete disease subtypes. Future work could extend
 499 the framework using deep clustering approaches combined with discrete salient representations,
 500 providing more control over subtype-conditioned synthetic data generation. This would allow more
 501 targeted simulations, enhance personalized modeling of disease effects, and provide a framework for
 502 exploring complex heterogeneity in neurological disorders. In summary, our framework provides a
 503 foundation for synthetic neuroimaging data generation that is practical, interpretable, and clinically
 504 relevant. By generating paired healthy and diseased samples, it enables personalized modeling of
 505 disease effects, supports subtype discovery, and facilitates the study of disease heterogeneity. With
 506 further extension to other multi-cohort datasets and target-informed regularization of the model,
 507 this approach has the potential to address challenges related to limited sample sizes, imbalanced
 508 datasets, and privacy concerns, while providing a valuable resource for research, education, and
 509 broader access to neuroimaging-derived data.

510 REFERENCES

512 Hemn Barzan Abdalla, Yulia Kumar, Jose Marchena, Stephany Guzman, Ardalan Awlla, Mehdi
 513 Gheisari, and Maryam Cheraghly. The future of artificial intelligence in the face of data scarcity.
 514 *Computers, Materials & Continua*, 84(1), 2025.

515 Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv*
 516 *preprint arXiv:1902.04601*, 2019.

518 Aidas Aglinskas, Joshua K Hartshorne, and Stefano Anzellotti. Contrastive machine learning reveals
 519 the structure of neuroanatomical variation within autism. *Science*, 376(6597):1070–1074, 2022.

521 Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, Ahmed Shihab Albahri, Bashar
 522 Sami Nayyef Al-Dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-
 523 Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges,
 524 solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.

525 André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle
 526 Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv*
 527 *preprint arXiv:2401.02524*, 2024.

529 Ganesh B Chand, Dominic B Dwyer, Guray Erus, Aristeidis Sotiras, Erdem Varol, Dhivya Srinivasan,
 530 Jimit Doshi, Raymond Pomponio, Alessandro Pignoni, Paola Dazzan, et al. Two distinct
 531 neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain*, 143(3):1027–
 532 1038, 2020.

533 Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning
 534 risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE*
 535 *International Conference on Data Mining (ICDM)*, pp. 787–792. IEEE, 2017.

537 Sai Spandana Chintapalli, Rongguang Wang, Zhijian Yang, Vasiliki Tassopoulou, Fanyang Yu,
 538 Vishnu Bashyam, Guray Erus, Pratik Chaudhari, Haochang Shou, and Christos Davatzikos. Gen-
 539 erative models of mri-derived neuroimaging features and associated dataset of 18,000 samples.
Scientific Data, 11(1):1330, 2024.

540 Andrew Kweku Conduah, Sebastian Ofoe, and Dorothy Siaw-Marfo. Data privacy in healthcare:
 541 Global challenges and solutions. *Digital Health*, 11:20552076251343959, 2025.
 542

543 Aoyan Dong, Nicolas Honnorat, Bilwaj Gaonkar, and Christos Davatzikos. Chimera: Clustering of
 544 heterogeneous disease effects via distribution matching of imaging patterns. *IEEE transactions*
 545 *on medical imaging*, 35(2):612–621, 2015.

546 Jimit Doshi, Guray Erus, Yangming Ou, Susan M Resnick, Ruben C Gur, Raquel E Gur, Theodore D
 547 Satterthwaite, Susan Furth, Christos Davatzikos, Alzheimer’s Neuroimaging Initiative, et al.
 548 Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and pa-
 549 rameters, and locally optimal atlas selection. *Neuroimage*, 127:186–195, 2016.
 550

551 Simon F Eskildsen, Pierrick Coupé, Vladimir S Fonov, Jens C Pruessner, D Louis Collins,
 552 Alzheimer’s Disease Neuroimaging Initiative, et al. Structural imaging biomarkers of alzheimer’s
 553 disease: predicting disease progression. *Neurobiology of aging*, 36:S23–S31, 2015.

554 Mohamad Habes, Raymond Pomponio, Haochang Shou, Jimit Doshi, Elizabeth Mamourian, Guray
 555 Erus, Ilya Nasrallah, Lenore J Launer, Tanweer Rashid, Murat Bilgel, et al. The brain chart of ag-
 556 ing: machine-learning analytics reveals links between brain aging, white matter disease, amyloid
 557 burden, and cognition in the istaging consortium of 10,216 harmonized mr scans. *Alzheimer’s &*
 558 *Dementia*, 17(1):89–102, 2021.
 559

560 HyeJeong Jeong, Jeong-Moo Lee, Hyeong Seok Kim, Hochang Chae, So Jeong Yoon, Sang Hyun
 561 Shin, In Woong Han, Jin Seok Heo, Ji Hye Min, Seung Hyup Hyun, et al. Synthetic data gener-
 562 ation method improves risk prediction model for early tumor recurrence after surgery in patients
 563 with pancreatic cancer. *Scientific Reports*, 15(1):31885, 2025.

564 Benjamin Lam, Mario Masellis, Morris Freedman, Donald T Stuss, and Sandra E Black. Clin-
 565 ical, imaging, and pathological heterogeneity of the alzheimer’s disease syndrome. *Alzheimer’s*
 566 *research & therapy*, 5(1):1, 2013.
 567

568 Robin Louiset, Edouard Duchesnay, Antoine Grigis, Benoit Dufumier, and Pietro Gori. Sep-
 569 vae: a contrastive vae to separate pathological patterns from healthy ones. *arXiv preprint*
 570 *arXiv:2307.06206*, 2023.

571 Robin Louiset, Edouard Duchesnay, Antoine Grigis, and Pietro Gori. Separating common from
 572 salient patterns with contrastive representation learning. *arXiv preprint arXiv:2402.11928*, 2024.
 573

574 Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-wasserstein distance
 575 between gaussian measures. *Information Geometry*, 5(1):289–323, 2022.

576 Andre F Marquand, Ilead Rezek, Jan Buitelaar, and Christian F Beckmann. Understanding het-
 577 erogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological*
 578 *psychiatry*, 80(7):552–561, 2016.

580 Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE*
 581 *international conference on data science and advanced analytics (DSAA)*, pp. 399–410. IEEE,
 582 2016.

583 Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for
 584 tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–
 585 869, 2020.

587 Vasileios C Pezoulas, Nikolaos S Tachos, George Gkois, Iacopo Olivotto, Fausto Barlocco, and
 588 Dimitrios I Fotiadis. Bayesian inference-based gaussian mixture models with optimal components
 589 estimation towards large-scale synthetic data generation for in silico clinical trials. *IEEE Open*
 590 *Journal of Engineering in Medicine and Biology*, 3:108–114, 2022.

592 Angelo Picardi, Cinzia Viroli, Lorenzo Tarsitani, Rossella Miglio, Giovanni de Girolamo, Giuseppe
 593 Dell’Acqua, and Massimo Biondi. Heterogeneity and symptom structure of schizophrenia. *Psy-
 chiatry research*, 198(3):386–394, 2012.

594 Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth
 595 Mamourian, Vishnu Bashyam, Ilya M Nasrallah, Theodore D Satterthwaite, Yong Fan, et al.
 596 Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the
 597 lifespan. *NeuroImage*, 208:116450, 2020.

598 Hang Qu, Haitao Ge, Liping Wang, Wei Wang, and Chunhong Hu. Volume changes of hippocampal
 599 and amygdala subfields in patients with mild cognitive impairment and alzheimer’s disease. *Acta*
 600 *Neurologica Belgica*, 123(4):1381–1393, 2023.

601 Jean-Francois Rajotte, Robert Bergen, David L Buckeridge, Khaled El Emam, Raymond Ng, and
 602 Elissa Strome. Synthetic data as an enabler for machine learning applications in medicine.
 603 *Iscience*, 25(11), 2022.

604 Adib Bin Rashid and MD Ashfakul Karim Kausik. Ai revolutionizing industries worldwide: A
 605 comprehensive overview of its diverse applications. *Hybrid Advances*, 7:100277, 2024.

606 Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent
 607 variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33,
 608 pp. 4862–4869, 2019.

609 Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep
 610 conditional generative models. *Advances in neural information processing systems*, 28, 2015.

611 Arthur W Toga and Karen L Crawford. The alzheimer’s disease neuroimaging initiative informatics
 612 core: a decade in review. *Alzheimer’s & Dementia*, 11(7):832–839, 2015.

613 Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. Moment matching deep contrastive latent
 614 variable models. *arXiv preprint arXiv:2202.10560*, 2022.

615 Jennifer L Whitwell. Progression of atrophy in alzheimer’s disease and related disorders. *Neurotox-
 616 icticity research*, 18(3):339–346, 2010.

617 Zhijian Yang, Ilya M Nasrallah, Haochang Shou, Junhao Wen, Jimit Doshi, Mohamad Habes, Guray
 618 Erus, Ahmed Abdulkadir, Susan M Resnick, Marilyn S Albert, et al. A deep learning framework
 619 identifies dimensional representations of alzheimer’s disease from brain structure. *Nature com-
 620 munications*, 12(1):7065, 2021.

621

A APPENDIX

A.1 ABLATION ANALYSIS OF THE CA FRAMEWORK AND MMD ALIGNMENT

621 To assess the utility of the contrastive autoencoder (CA) structure and the MMD alignment term, we
 622 performed an ablation study comparing several model variants: 1) Conditional VAE (cVAE)(Sohn
 623 et al., 2015) – a baseline without the contrastive structure, included to evaluate the added value of
 624 CA. 2) CA-VAE with $\lambda = 0$ – the CA structure without any MMD alignment, i.e., no decoupling of
 625 the shared latent space. 3) CA-VAE with $\lambda = 5$, $\lambda = 10$, and $\lambda = 15$ – the CA structure with increasing
 626 MMD regularization strength. 5) CA-VAE with Total Correlation (TC) regularizer – replacing
 627 MMD with the TC loss described in the original contrastive VAE paper (Abid & Zou, 2019). We
 628 evaluated fidelity of synthesized samples using 2-Wasserstein distances (lower is better for fidelity;
 629 for background-target separation, values closer to the real separation are better). Bootstrap resam-
 630 pling ($N = 1000$) was used to compute 95% confidence intervals. All models were trained on the
 631 semi-synthetic data. Table A1 shows that Contarstiev VAE framework with MMD alignment penalty
 632 helps generate synthetic background and target samples with higher fidelity.

633 To assess whether the shared latent representation was successfully decoupled (i.e. Equation 6)
 634 under different configurations of the CA framework, we evaluated how well a simple logistic
 635 regression classifier could distinguish target from background samples using only the common latent
 636 means produced by the encoder q_{ϕ_z} . The latent means provide the deterministic component of the
 637 posterior and are standard for assessing latent-space structure. The classifier was trained using 5-
 638 fold cross-validation on latent codes extracted from the semi-synthetic dataset. An AUC near 0.5
 639 indicates random classification, meaning that the latent space cannot separate the two datasets and

648

649

Table A1: Ablation study of MMD alignment in semi-synthetic experiment

Model Variant	Generated vs Real Target	Generated vs Real Background	Unpaired Background vs Target
Semi-Synthetic data	-	-	20.36
Conditional VAE	59.45 (58.62-62.45)	61.7 (60.79-64.66)	12.48 (11.9-14.94)
Contrastive VAE, $\lambda = 0$	62.18 (60.94-65.29)	51.33 (50.73-55.05)	14.99 (14.4-16.81)
Contrastive VAE, $\lambda = 5$	59.98 (59.28-62.82)	53.19 (52.25-57.26)	15.01 (14.28-17.08)
Contrastive VAE, $\lambda = 10$	60.47 (59.73-63.15)	51.83 (51.42-55.06)	15.16 (14.48-17.35)
Contrastive VAE, $\lambda = 15$	62.05 (60.86-64.91)	54.21 (52.99-58.43)	14.86 (14.15-16.91)
Contrastive VAE, TC loss	61.14 (60.2-64.14)	56.46 (55.34-60.9)	13.58 (12.9-15.75)

650

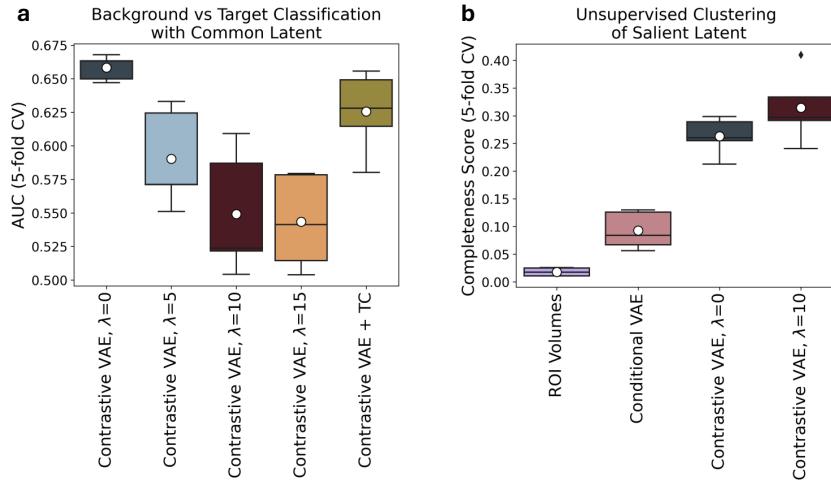
651

652

653

654

655



656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

Figure A1: Supplementary results from semi-synthetic experiment: a) Boxplot of AUC scores from a logistic regression classifier trained to distinguish target from background samples using the common latent means. Lower AUC values indicate better latent-space decoupling across different Contrastive VAE configurations. b) Clustering performance of salient latent representations in the semi-synthetic dataset. K-means clustering was applied with $k=4$ (three disease subtypes plus background), and clustering accuracy was quantified using the completeness score across 5-fold cross-validation. Higher scores indicate better recovery of subtype-specific variation.

672

673

674

675

676

677

678

679

680

681

682

is therefore properly decoupled. Notably, when strength of alignment penalty λ increases to 10 the mean AUC drops to 0.549, indicating improved decoupling (Figure A1a).

To quantitatively evaluate whether the latent space captures disease heterogeneity, we performed K-means clustering (with $K=4$, reflecting the three target subtypes plus the background in semi-synthetic data) on the salient latent means inferred from the semi-synthetic dataset. Clustering performance was assessed using the completeness score, with clustering repeated across 5-fold cross-validation to ensure robustness. The CA-VAE with MMD $\lambda = 10$ achieved the highest mean completeness score (0.315), outperforming both the raw ROI volumes and the latent space of the conditional VAE (Figure A1b). This indicates that contrastive analysis along with MMD alignment in the common latent space not only enforces decoupling but also enables the salient latent space to encode meaningful subtype-specific variation.

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

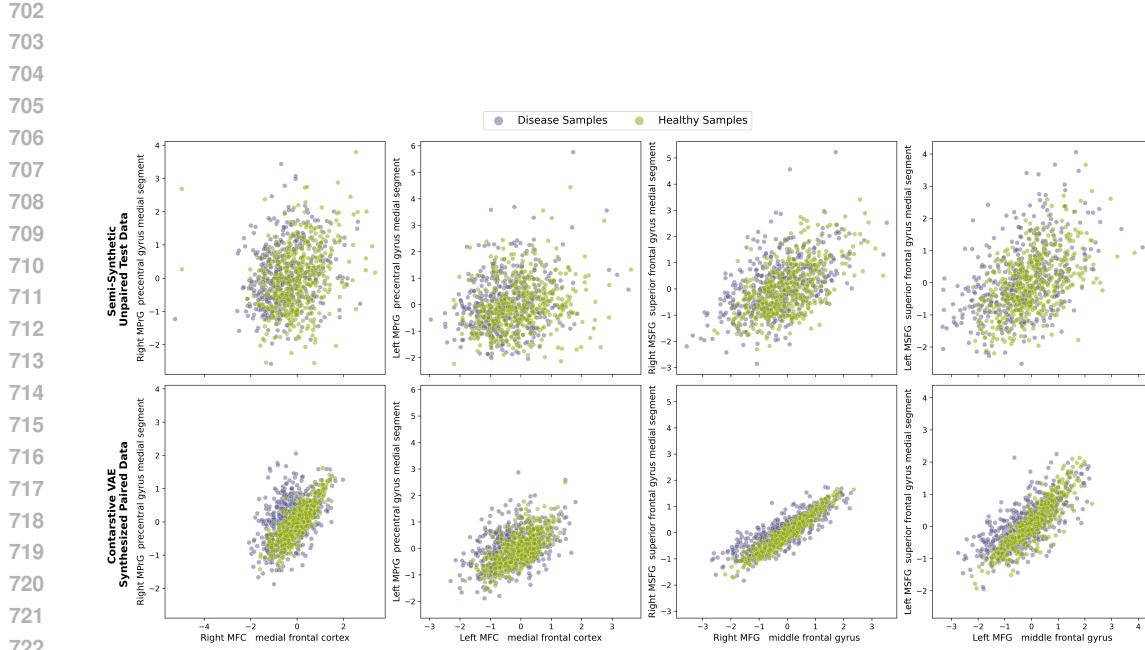


Figure A2: Supplementary results from semi-synthetic experiment: Distribution of standardized brain volumes in the semi-synthetic test data (top row) and the contrastive VAE-generated paired dataset (bottom row).

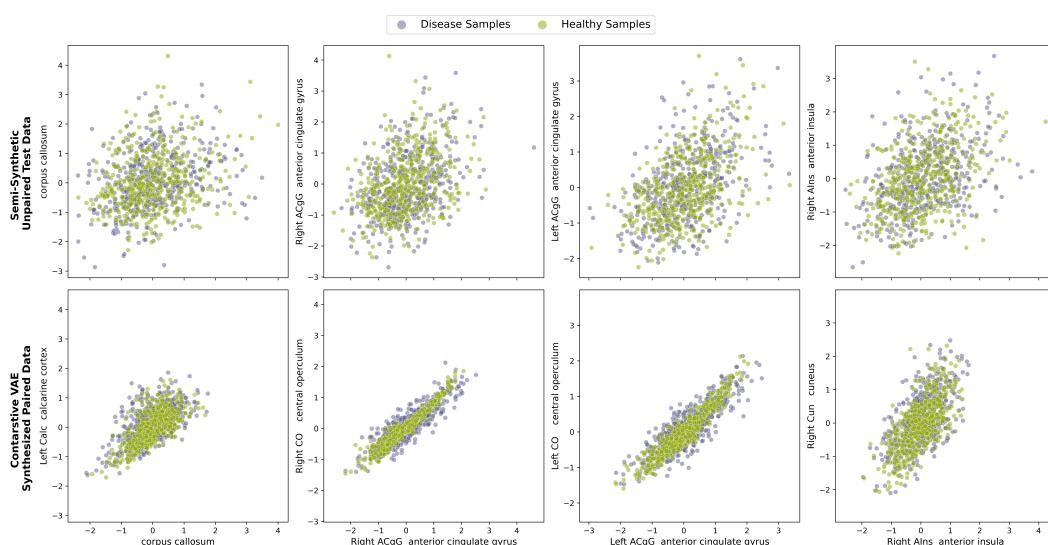
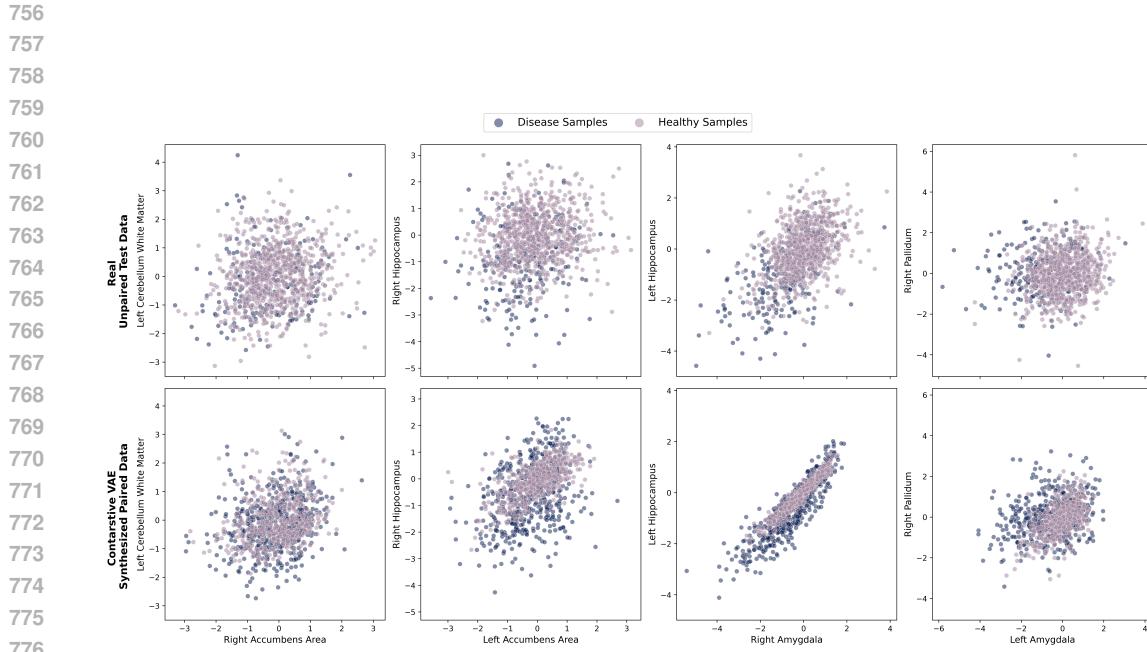
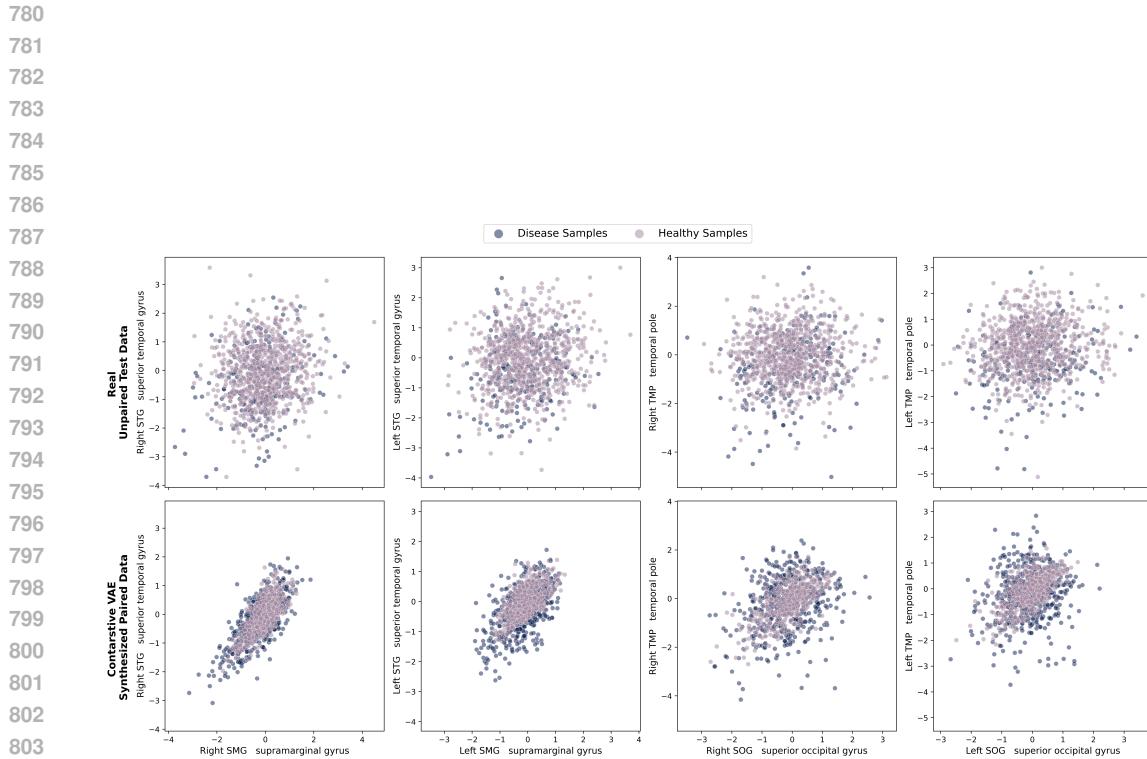


Figure A3: Supplementary results from semi-synthetic experiment: Distribution of standardized brain volumes in the semi-synthetic test data (top row) and the contrastive VAE-generated paired dataset (bottom row).



777 Figure A4: Supplementary results from real data experiment: Distribution of standardized brain
778
779 volume in the real test data (top row) and the contrastive VAE-generated paired dataset (bottom
row).



804 Figure A5: Supplementary results from real data experiment: Distribution of standardized brain
805 volume in the real test data (top row) and the contrastive VAE-generated paired dataset (bottom
row).

806
807
808
809