The Anatomy of Synthesis: Simulating Changes in the Human Brain over Time through Diffeomorphic Deformations

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Anirudh Kaushik
2020111015
anirudh.ka@research.iiit.ac.in

Advisor: Professor Jayanthi Sivaswamy

International Institute of Information Technology, Hyderabad
500 032, India

June 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***The Anatomy of Synthesis: Simulating Changes in the Human Brain over Time through Diffeomorphic Deformations*** by *Anirudh Kaushik* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Professor Jayanthi Sivaswamy

Acknowledgements

I would like to express my deepest gratitude to my research advisor, Dr. Jayanthi Sivaswamy. Her unwavering support, insightful guidance, and constructive feedback have helped me throughout my research journey. I am fortunate to benefit from her dedication to my academic success, and her personal support to help me work at my pace.

I would also like to extend my appreciation to my co-author, Naren Akash. His collaborative spirit, and willingness to share his knowledge has been essential to the development of this research. The insightful discussions and collaborative efforts we shared have significantly improved the quality of this work.

I am also grateful to my parents for their constant support and encouragement throughout my academic pursuits. Their belief in my abilities has been a source of strength and motivation.

My gratitude extends to my lab mates, Alphin, Abhishek, Rudransh, Soham, Samruddhi and Shreyas. Their presence in the lab fostered a collaborative environment. Their willingness to lend a hand with experiments, engage in discussions, and offer insights during troubleshooting sessions proved invaluable throughout this research journey.

Abstract

The human brain undergoes continuous structural changes throughout the lifespan, driven by a complex interplay of aging processes, environmental influences, and disease-related mechanisms. Patterns of structural change—particularly atrophy associated with tissue loss and shrinkage—emerge gradually over time and are observable using medical imaging techniques. While these changes are shaped by common biological mechanisms, they are also highly individualized, influenced by factors such as lifestyle, and neurological conditions like Alzheimer's Disease (AD), Parkinson's disease, tumors, and stroke. Understanding the progression of these changes—both at the individual level and across populations—is critical for advancing our knowledge of healthy aging and the dynamics of neurodegenerative disease.

To study how brain structure evolves over time, researchers rely on longitudinal neuroimaging: repeated imaging of the same individuals at multiple timepoints. Unlike cross-sectional imaging, which captures a single snapshot per subject, longitudinal scans provide a temporal sequence that enables direct observation of anatomical trajectories. These sequences allow for the measurement of rates of change, identification of early biomarkers, and modeling of disease progression in a subject-specific manner.

However, acquiring complete longitudinal datasets in practice remains challenging. Subject dropout, missed clinical visits, and protocol variability often result in missing scans, interrupting the temporal continuity required for accurate modeling. These gaps limit the effectiveness of methods that rely on temporally complete inputs and can bias downstream analyses. Imputing the missing scan to complete the subject's imaging timeline is therefore a critical step toward enabling robust longitudinal modeling and improving our understanding of neurodegenerative processes.

This thesis addresses the challenges of modeling and analyzing longitudinal brain changes by developing anatomically grounded methods for data imputation, latent space disentanglement, and downstream trajectory analysis. We first introduce SynBADD, a deformation-based framework that synthesizes missing brain scans by predicting physiologically plausible stationary velocity fields (SVFs)—parametric fields that encode smooth, invertible deformations over space and time—rather than directly generating full image intensities. By operating in the deformation space, SynBADD preserves anatomical coherence and spatial fidelity while mitigating the artifacts typically associated with intensity-based synthesis approaches.

Building on this foundation, we propose DIVA, a metadata-informed variational autoencoder designed to learn a temporally disentangled latent space for modeling brain morphological changes. DIVA is trained on synthetically augmented Stationary Velocity Fields (SVFs), enriching intra-subject variation and improving the model's capacity to generalize across limited real-world samples. Age, disease label, and temporal

information are explicitly disentangled through conditional bottleneck supervision, allowing the learned latent representations to reflect meaningful clinical and chronological factors. This disentanglement enhances temporal predictability, supports more accurate subject-specific trajectory modeling, and enables the use of powerful transformer-based architectures for latent space interpolation and prediction.

We further extend our analysis in an "Analysis by Synthesis" framework, investigating how metadata-informed conditioning affects generation quality and how different temporal reasoning strategies (past-only, future-only, and bidirectional) impact anatomical plausibility. We perform trajectory-based analyses of generated scans, evaluating how well imputed data aligns with true subject-specific anatomical trends across key regions of interest (ROIs) such as the hippocampus and parahippocampus. Additionally, we assess subgroup differences by age, sex, and disease status, and evaluate downstream task performance with and without synthetic imputation.

Through comprehensive evaluations on the Alzheimer's Disease Neuroimaging Initiative (ADNI) longitudinal dataset, our approaches achieve substantial improvements over state-of-the-art baselines in both image fidelity and clinically relevant anatomical accuracy. Beyond technical advancements, this work provides new insights into the modeling of individualized brain aging patterns and opens pathways for data augmentation in clinical studies where longitudinal completeness is often unattainable.

Contents

List of Figures

List of Tables

Abbreviations

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AE-GAN	Auto-Encoding GAN for 3D Brain MRI Generation
DDPMs	Denoising diffusion probabilistic models
HA-GAN	Hierarchical Amortized GAN for 3D High-Resolution Medical Image Synthesis
LDMs	Latent Diffusion Models
Red-GAN	Red-GAN: Attacking Class Imbalance via Conditioned Generation
SVFs	Stationary Velocity Fields
VAE	Variational Autoencoder

Symbols

ϕ	stationary velocity field deformation
$J_\phi(x)$	Jacobian matrix of the deformation field ϕ at point x
$\det(J_\phi(x))$	Determinant of the Jacobian matrix, representing local volume change
$\phi_{T_i \to 0}$	stationary velocity field deformation from time T_i to baseline T_0
z_i	latent code representing deformation at time T_i
s_i	skip features extracted from SVF at time T_i
$\bar{s}$	aggregated spatial prior from available skip features
$\hat{z}_p$	predicted latent code at missing timepoint T_p
$\{z_j\}_{j \neq p}$	set of observed latent codes excluding missing timepoint
$\tilde{\phi}_{T_p \to 0}$	intermediate decoded SVF prior to refinement
$\hat{\phi}_{T_p \to 0}$	refined SVF after U-Net deformation generator

List of Related Publications

[P1] Anirudh Kaushik, Naren Akash RJ and Jayanthi Sivaswamy, **"SynBADD: Longitudinal Medical Image Imputation Via Synthetic Bidirectional Anatomy-guided Diffeomorphic Deformations"**, in review at *Medical Image Analysis*, vol. XX, pp. XXX–XXX, 2025.

Related co-author publications:

[P2] Naren Akash R J, Anirudh Kaushik, and Jayanthi Sivaswamy, **"Evidence-Driven Differential Diagnosis of Malignant Melanoma"**, in proceedings of *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

Chapter 1

Introduction

1.1 Motivation for Brain Volume Imputation

According to the World Health Organization (WHO) [1], someone in the world develops dementia every three seconds. As of 2019, approximately 55 million people worldwide were living with dementia, a number projected to increase to 139 million by 2050 [2]. Alzheimer's Disease (AD), the most common form of dementia, accounts for 60–70% of all cases and is characterized by progressive cognitive decline driven by underlying neurodegenerative processes [3]. Importantly, up to 40% of dementia cases could potentially be prevented or delayed with early detection and intervention [4].

Longitudinal neuroimaging data are critical for studying the onset and progression of neurodegenerative diseases such as AD. Unlike cross-sectional datasets, which are confounded by inter-subject variability, longitudinal imaging allows for the direct observation of within-subject anatomical changes over time, providing a more precise understanding of disease trajectories [5]. Such data enable the study of rates of change, identification of early biomarkers, and temporal modeling of structural decline relative to disease onset.

Despite their importance, longitudinal neuroimaging datasets often suffer from missing data. Even in well-established cohorts such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), patient dropout, missed visits, and protocol inconsistencies introduce substantial gaps [6]. These discontinuities degrade the performance of computational models that rely on temporally continuous inputs for disease progression modeling, prognosis, and treatment planning [7–10].

Addressing missing data in longitudinal imaging studies is critical for improving the fidelity of disease trajectory modeling, enhancing biomarker discovery, and supporting clinical decision-making. Imputing missing brain volumes allows for better temporal continuity, more accurate interpolation of anatomical changes, and improved modeling of gradual disease processes such as neurodegeneration, tumor progression, and organ development.

Recent approaches to longitudinal data imputation have predominantly leveraged generative models such as generative adversarial networks (GANs) and diffusion models. While diffusion-based methods have

demonstrated impressive image synthesis quality in medical imaging applications, generating full 3D brain volumes from scratch remains computationally expensive and prone to artifacts that compromise anatomical consistency [11–13]. These limitations reduce the clinical utility of intensity-based synthesis, particularly for applications requiring high-fidelity structural modeling. More recently, physics-inspired approaches that model anatomical deformations, rather than raw image intensities, have emerged as promising alternatives for anatomically grounded image generation.

1.2 Challenges in Medical Image Generation

Medical image generation, particularly in the longitudinal setting, remains a complex and underexplored problem. Existing methods for sequence-aware brain volume generation, such as Sequence Aware Diffusion Model (SADM) [14], predominantly focus on predicting future states from prior scans, often overlooking the valuable information embedded in subsequent observations. This unidirectional modeling limits the reconstruction of temporally coherent subject-specific trajectories, especially when anatomical changes are nonlinear and vary widely across individuals.

The high computational demands of synthesizing full 3D brain volumes further complicate the generation process. Producing volumetric data directly in the intensity space requires substantial memory and computational resources, restricting scalability and practical deployment in clinical settings. Diffusion-based generative models, while producing high-quality outputs, are particularly expensive at inference time, requiring approximately 13 seconds per slice and around **4 minutes** to synthesize a full $128 \times 128 \times 128$ volume. Models that generate voxel intensities from scratch are also prone to spatial inconsistencies and often introduce anatomically implausible artifacts, particularly when trained on small datasets. Some methods, such as Recovery of Missing Neuroimaging using Diffusion Models with Application to Alzheimer's Disease (ReMiND) [15], attempt to mitigate these challenges by synthesizing 2D slices and stacking them into 3D volumes; however, this approach disrupts volumetric continuity and leads to spatial artifacts across slices. The lack of explicit anatomical constraints in intensity-based generation fundamentally limits the structural fidelity and clinical validity of the synthesized outputs.

These issues are compounded by the irregular and sparse nature of real-world longitudinal imaging datasets. Incomplete and uneven sampling across timepoints interrupts the continuity necessary for learning consistent spatiotemporal patterns. Missing scans, particularly during critical stages of disease progression, impair a model's ability to infer meaningful subject-specific anatomical trajectories and introduce biases in downstream predictive tasks.

Recent advances such as Nodeo: A neural ordinary differential equation based optimization framework for deformable image registration (NODEO) [16] and Neural Ordinary Differential Equations (Neural ODE) [17] have introduced frameworks for modeling temporal anatomical changes through image regression, enabling the estimation of missing scans by sampling intermediate timepoints along a continuous trajectory. While these approaches represent a significant step forward in longitudinal modeling, methods

such as Image Sequence Regression Based on Neural Ordinary Differential Equations (NODER) [18] still face practical limitations. In particular, they often require training separate models for each subject and each timepoint, which restricts scalability to larger cohorts and increases susceptibility to interpolation errors when estimating missing data points.

In summary, the generation of anatomically plausible and temporally consistent intermediate scans remains a significant challenge. Overcoming these limitations requires frameworks capable of bidirectional temporal reasoning, operating within anatomically constrained manifolds, and robustly handling sparse and irregular longitudinal data while preserving high spatial fidelity.

1.3 Methods for Synthetic Brain MRI Generation

The generation of synthetic brain MRI volumes has emerged as an important direction in medical imaging research, driven by the need to augment limited datasets, address class imbalance, and handle missing or incomplete data through imputation techniques. Unlike natural image generation tasks, brain MRI synthesis must account for strict anatomical and physiological constraints to ensure clinical plausibility. Early efforts, such as Hierarchical Amortized GAN for 3D High-Resolution Medical Image Synthesis (HA-GAN) [19], demonstrated the feasibility of generating high-resolution 3D thorax CT and brain MRI volumes using hierarchical generative adversarial network (GAN) architectures. Similarly, Auto-Encoding GAN for 3D Brain MRI Generation (AE-GAN) [20] combined variational autoencoders (VAEs) with GANs to mitigate two of the most common challenges faced by generative models: mode collapse in GANs and output blurriness in VAEs. These approaches established foundational techniques for synthesizing volumetric medical images, but did not yet introduce structured anatomical control over the generated content.

Subsequent developments in brain MRI synthesis introduced conditional semantic image synthesis as a means of introducing greater control over the generation process. Semantic synthesis inverts the traditional task of segmentation by generating MR images from segmentation masks rather than extracting masks from images. This approach enables direct control over anatomical attributes, allowing models to generate outputs with specified spatial configurations. For instance, Red-GAN: Attacking Class Imbalance via Conditioned Generation (Red-GAN) [21] employed the SPADE module to condition image synthesis on both pixel-level segmentation masks and global acquisition parameters, addressing challenges related to class imbalance and pathological variability. More recently, A Deep Generative Framework for Paired 3D Image and Mask Generation (Med-DDPM) [22] demonstrated the use of conditional diffusion models to synthesize high-resolution 3D brain MRIs, showing that diffusion-based methods can produce detailed and anatomically realistic volumes when guided by semantic information.

While semantic synthesis has enabled the generation of both healthy and pathological brain images, much of the existing work has concentrated on pathological cases, particularly the modeling of brain tumors. This focus stems from the ability to manipulate tumor-specific attributes—such as location, size, and morphology—while preserving the global brain structure. Studies such as ”Generating 3D Brain Tumor Regions

in MRI Using Vector-Quantization GAN" [23], "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition" [24], and "Automatic Brain Tumor Grading from MRI Data Using CNNs and Quality Assessment" [25] have demonstrated that focusing generation efforts on tumor regions, rather than entire brain volumes, can improve computational efficiency and enhance model training effectiveness by reducing redundancy.

Beyond semantic synthesis, paired generation frameworks have also been proposed to further align anatomical structure and image synthesis. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis (MedGen3D) [26] introduced a two-stage approach in which anatomical masks are generated first, followed by the synthesis of MR images conditioned on these masks. This design ensures that the structural layout of the generated image aligns closely with the underlying anatomy. However, the advantages of explicitly generating segmentation masks prior to image synthesis remain unclear, as no ablation studies have been conducted to compare this approach against post-hoc segmentation strategies.

More recently, the field has moved toward learning contrast-agnostic anatomical representations to enhance generalizability across different imaging protocols and acquisition settings. Contrast-Agnostic Anatomical Representations for Brain Imaging (Brain-ID) [27] introduced a model trained entirely on synthetic datasets augmented through mild-to-severe intra-subject perturbations. Brain-ID learns subject-specific anatomical embeddings that remain robust to variations in contrast, resolution, and imaging artifacts. Once trained, the model can be adapted to a variety of downstream tasks—including brain segmentation, super-resolution, and bias field correction—with minimal additional supervision. Such developments represent an important step toward creating clinically adaptable models capable of operating reliably across heterogeneous neuroimaging datasets.

1.4 Structural Brain Changes Across the Aging and Disease Spectrum

The human brain undergoes continuous structural changes throughout the lifespan, even in the absence of overt pathology. Normal aging is associated with gradual alterations in brain morphology, including reductions in brain volume, cortical thinning, and ventricular enlargement. These changes unfold over decades and, while measurable, generally have minimal impact on daily cognitive function in healthy individuals. Understanding the baseline trajectory of structural brain evolution is essential for contextualizing deviations that may signal the onset of pathological processes.

Among neurodegenerative disorders, Alzheimer's disease (AD) is the most prominent cause of accelerated brain atrophy, characterized by widespread cortical thinning, hippocampal volume loss, and ventricular expansion. Distinguishing between the expected anatomical changes of healthy aging and the pathological degeneration associated with AD is critical for early diagnosis and effective intervention. Cross-sectional imaging alone often fails to capture these subtle differences, underscoring the importance of longitudinal studies that can track structural trajectories within individuals over time.

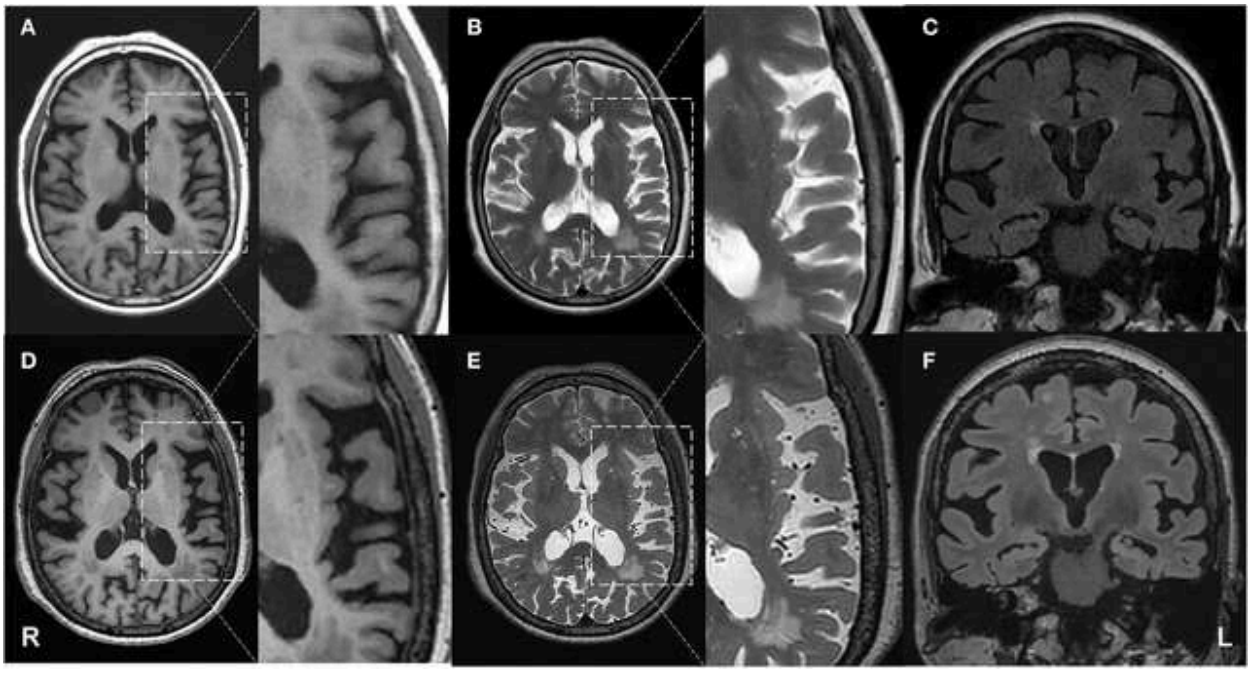


Figure 1.1: Progressive brain atrophy observed in structural MRI scans of a single subject over a two-year period. Top row: scans from February 15, 2019; bottom row: scans from April 25, 2021. (A, D) Axial T1-weighted images; (B, E) axial T2-weighted images; (C, F) coronal FLAIR images. Notable atrophy is observed in the left temporal lobe over time.

This section summarizes key findings regarding brain morphological changes across normal aging, mild cognitive impairment (MCI), and Alzheimer's disease, highlighting the essential role of longitudinal modeling in characterizing disease-specific progression patterns.

1.4.1 Brain Morphological Changes During Normal Aging

Brain atrophy—the gradual loss of brain tissue—is a fundamental aspect of the normal aging process. This decline begins as early as 30 to 40 years of age, well before clinical signs of cognitive impairment typically emerge. Figure 1.1 illustrates longitudinal brain atrophy in a single subject over a two-year period, with visible reductions in cortical volume [28].

Longitudinal studies estimate that, on average, whole-brain atrophy progresses at a relatively slow rate of approximately 0.32% per year throughout adulthood [29]. Among specific regions, the hippocampus—a structure critically involved in memory—exhibits volume loss at a rate of 1–2% annually [30]. Cortical thinning is also observed, although it occurs heterogeneously, with the prefrontal and parietal cortices demonstrating greater susceptibility than other areas [31]. While these anatomical changes are measurable, they typically do not impair daily cognitive functioning in healthy older adults, highlighting the distinction between normative aging and pathological neurodegeneration.

However, in some individuals, the rate and distribution of atrophy deviate from these normative patterns, marking the earliest structural signs of pathological processes such as mild cognitive impairment and Alzheimer's disease.

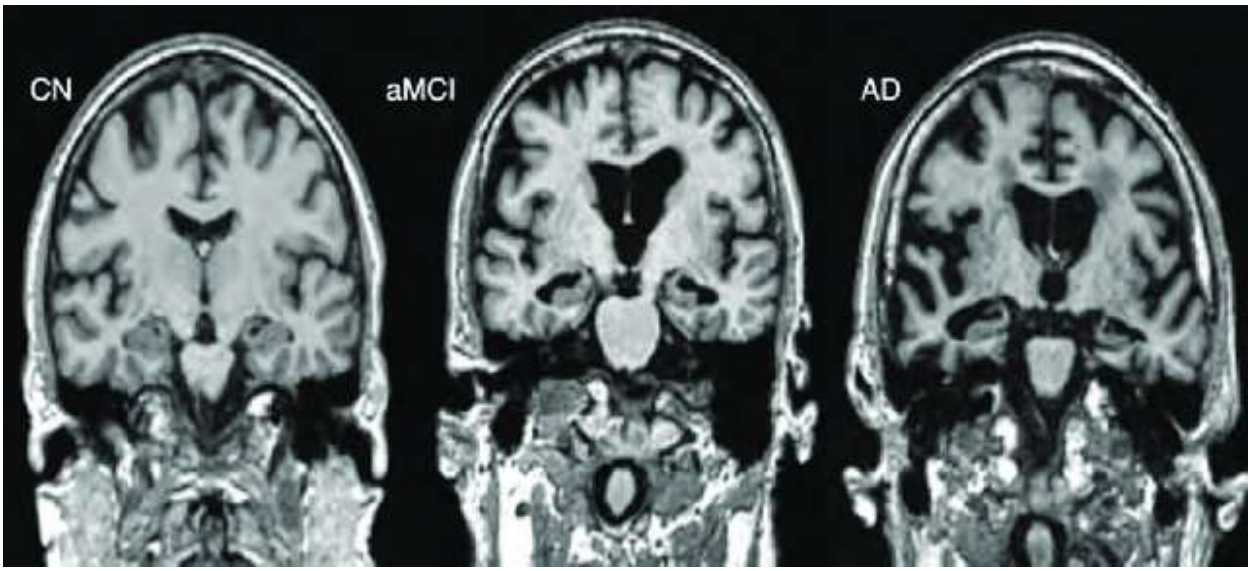


Figure 1.2: Representative structural MRI slices demonstrating progressive brain atrophy across clinical stages. Minimal atrophy is observed in cognitively normal (CN) individuals, which increases in amnestic mild cognitive impairment (aMCI) and becomes extensive in Alzheimer's disease (AD). Ventricular expansion and cortical thinning are visibly more pronounced with disease progression.

1.4.2 Structural Alterations in Mild Cognitive Impairment

Mild cognitive impairment (MCI) represents a transitional stage between normal aging and dementia, characterized by more pronounced anatomical degeneration. Individuals with MCI exhibit hippocampal atrophy rates that are intermediate between normal aging and Alzheimer's disease, estimated at 2–3% per year [32]. Cortical thinning becomes more noticeable, particularly in the hippocampus, parahippocampal gyrus, and temporal lobes [33], regions critical for memory and navigation. Although individuals with MCI maintain relatively preserved functional independence, these structural changes are predictive of elevated risk for progression to AD.

1.4.3 Accelerated Degeneration in Alzheimer's Disease

Alzheimer's disease (AD) is associated with accelerated and widespread brain atrophy compared to normal aging. Hippocampal volume loss in AD can reach 4–5% per year [34], and cortical thinning is severe and extensive, affecting temporal, parietal, and frontal regions [35]. Ventricular enlargement, particularly the expansion of the lateral ventricles, is another hallmark of AD progression, reflecting the loss of surrounding brain tissue [36]. These morphological changes are strongly correlated with cognitive decline, functional impairment, and disease staging. An example of this progression is illustrated in Figure 1.2, which compares brain scans of individuals with normal cognition, amnestic mild cognitive impairment (aMCI), and Alzheimer's disease [37].

1.4.4 Distinguishing Normal Aging from Alzheimer's Disease with Machine Learning and Biomechanical Models

Recent advances in artificial intelligence and biomechanical modeling have enabled more nuanced differentiation between healthy aging and Alzheimer's-related degeneration. Machine learning approaches can distinguish typical aging patterns from pathological changes by analyzing subtle differences in brain morphology [38]. Biomechanical simulations have also been employed to model how neurodegeneration alters brain structure over time, providing predictive insights into dementia risk [39]. Such approaches underscore the importance of longitudinal data in capturing disease-specific trajectories rather than relying solely on static snapshots.

Longitudinal neuroimaging studies have established that brain atrophy progresses gradually during normal aging, typically affecting the frontal and parietal cortices at a relatively slow rate of 0.3–0.5% per year [40, 41]. In contrast, mild cognitive impairment (MCI) is associated with more pronounced degeneration, particularly in the hippocampus and entorhinal cortex, where volume loss accelerates to 1–2% annually [42, 43]. As Alzheimer's disease (AD) develops, this atrophy becomes more rapid and spatially widespread, with regions such as the temporal and parietal lobes exhibiting significant shrinkage and the hippocampus losing up to 4–5% of its volume per year [44, 45]. Additionally, ventricular enlargement—especially in the lateral ventricles—serves as a hallmark of AD progression [46]. These changes are accompanied by severe cortical thinning and extensive neuronal loss across multiple regions. A comparative summary of these morphological trajectories across normal aging, MCI, and AD is provided in Table 1.1

1.5 Motivation for Longitudinal Modeling

The progression from normal aging to Alzheimer's disease follows a continuum of gradual anatomical changes that unfold over years or decades. Cross-sectional imaging cannot fully capture the dynamics of this evolution, as it conflates inter-subject variability with true longitudinal trends. In contrast, longitudinal modeling enables the detection of subject-specific trajectories, early disease signatures, and the timing of structural inflection points that signal pathological progression. Therefore, accurately capturing brain volume changes over time is essential for advancing early diagnosis, monitoring disease progression, and evaluating interventions in neurodegenerative disorders.

1.6 SVFs for Modelling Morphological Changes in the Brain

Stationary velocity fields (SVFs) provide a mathematically grounded way of generating realistic spatial transformations between medical images while ensuring properties such as smoothness, invertibility, and topology preservation.

What is an SVF? At a high level, an SVF defines a continuous field of vectors over space that describes the velocity at which each point in the image domain moves. Unlike general time-varying velocity fields, a stationary velocity field is constant with respect to time—hence the term "stationary." The velocity assigned to each point does not change as we move along the transformation trajectory. This assumption allows SVFs

Feature	Normal Aging	Mild Cognitive Impairment (MCI)	Alzheimer's Disease (AD)
Atrophy Rate	Slow (0.3–0.5%/year)	Moderate (1–2%/year)	Fast (2–3%/year)
Earliest Affected Regions	Frontal & Parietal Cortex (subtle)	Hippocampus, Entorhinal Cortex	Hippocampus, Temporal & Parietal Lobes
Ventricular Enlargement	Mild	Moderate	Significant
Cortical Thinning	Gradual (0.2mm/decade)	Noticeable in key regions	Severe, widespread
Hippocampal Volume Loss	1–2% per year	2–3% per year	4–5% per year

Table 1.1: **Comparative summary of morphological brain changes across Normal Aging, Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD).** Key features of neurodegeneration are shown across clinical stages, including rates of atrophy, cortical thinning, and hippocampal volume loss. While gradual structural decline is expected in normal aging, MCI and AD demonstrate increasingly rapid and regionally specific degeneration. The red-highlighted column emphasizes the accelerated progression and severity of changes associated with AD.

to model complex deformations while remaining computationally efficient and mathematically tractable. Figure 1.3 illustrates an example of an SVF across axial, coronal, and sagittal planes, while Figure 1.4 shows how an SVF-defined deformation field warps a regular grid and transforms a moving image toward alignment with a fixed reference [47].

Mathematical Formulation. Let $\phi : \Omega \to \mathbb{R}^3$ denote a diffeomorphic transformation of the image domain $\Omega \subset \mathbb{R}^3$. In the SVF framework, we generate this transformation by integrating a velocity field $v : \Omega \to \mathbb{R}^3$ over unit time via the exponential map:

$$\phi = \exp(v),$$

where $\exp(v)$ denotes the integration of the ordinary differential equation (ODE):

$$\frac{d\phi_t(x)}{dt} = v(\phi_t(x)), \quad \phi_0(x) = x,$$

evaluated at $t = 1$. Since v is assumed to be stationary (i.e., time-invariant), the resulting transformation is diffeomorphic—meaning it is smooth, invertible, and preserves local topology.

Why "Stationary"? The term"stationary" does not imply that the underlying anatomy is static, but rather that the velocity field used to define the transformation is fixed in time. The anatomy evolves over time by flowing along this fixed field. This yields a consistent and smooth deformation path that can model changes such as brain atrophy or tissue shrinkage.

SVFs as a Generative Framework. In the context of image generation, SVFs serve as a compact and interpretable generative representation. Instead of directly synthesizing pixel or voxel intensities, we generate a deformation field that is applied to an existing source image (e.g., a baseline brain scan). This ensures anatomical coherence because the generated image inherits its structural consistency from the source while being morphed through a smooth, plausible transformation. In this sense, SVFs allow the synthesis of new anatomical states without hallucinating entirely new intensity patterns, which is particularly advantageous in medical domains with limited data and strict physiological constraints.

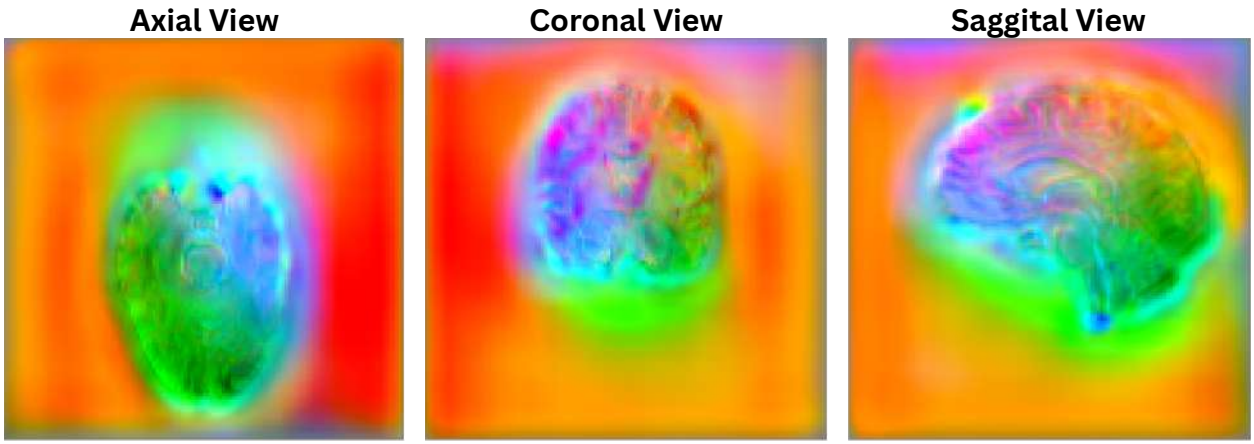


Figure 1.3: Visualization of a stationary velocity field (SVF) overlaid on axial, coronal, and sagittal views. The SVF is encoded as a three-channel displacement field, where each channel represents spatial velocity along one of the x, y, or z axes. For visualization purposes, magnitudes are rescaled and mapped to RGB color space to convey directional flow and intensity.

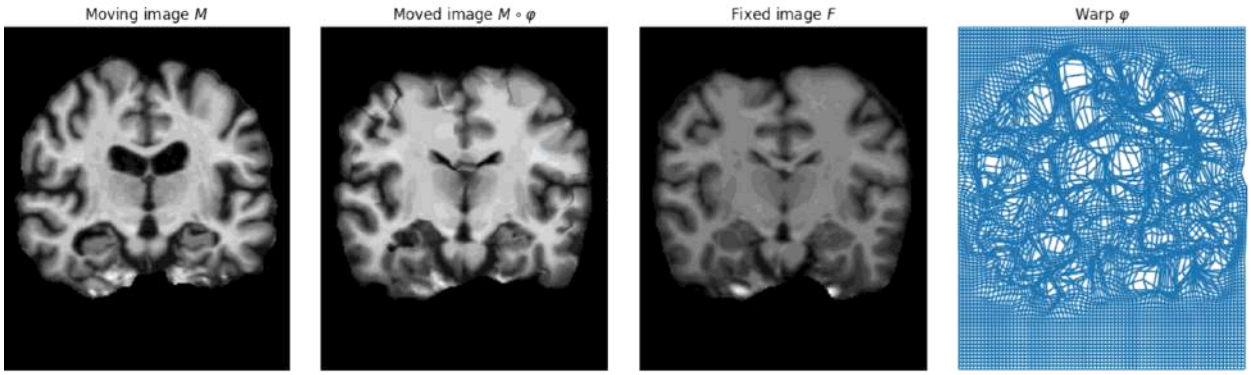


Figure 1.4: Demonstration of SVF-driven deformation. The left panel shows the original moving image, which is warped using the predicted deformation field ϕ to align with a fixed reference image. The rightmost panel shows a regularly spaced grid deformed by ϕ, illustrating the smooth and continuous nature of the transformation. Image adapted from [47].

Relation to Large Deformation Models. The SVF model is related to the broader family of diffeomorphic registration methods, such as Large Deformation Diffeomorphic Metric Mapping (LDDMM). However, SVFs simplify the computational burden by assuming stationarity, enabling efficient optimization and learning while still preserving diffeomorphic properties.

This thesis demonstrates the potential of SVFs as a generative backbone for missing scan imputation and trajectory modeling. By predicting subject-specific SVFs in a learned latent space, the model can synthesize anatomically plausible future or intermediate scans by warping a known baseline image.

1.7 Image Imputation and Disease Simulation

Longitudinal imaging datasets often suffer from missing scans due to subject dropout, inconsistent scanning protocols, or logistical constraints. Missing data poses a significant challenge for downstream modeling tasks that rely on temporally coherent anatomical trajectories. To address this, recent research has focused on two complementary goals: (i) imputation of missing images to restore continuity in longitudinal sequences, and (ii) simulation of aging and disease progression to model morphological changes over time. This section reviews major developments in these areas, highlighting how deep generative models have been adapted for longitudinal modeling.

1.7.1 Image Imputation for Longitudinal Continuity

Early approaches to image imputation emphasized generating missing intermediate scans to restore temporal continuity. One line of work introduced diffusion-based deformation models, where Denoising diffusion probabilistic models (DDPMs) were combined with learned deformation fields to generate intermediate timepoints. Diffusion Deformable Model for 4D Temporal Medical Image Generation [48] exemplifies this approach, synthesizing continuous temporal frames between source and target medical images while preserving anatomical topology. Originally demonstrated on 4D cardiac MRI data, such frameworks offer a promising avenue for interpolating brain anatomical evolution over time.

Further advancing sequence modeling, the Sequence-Aware Diffusion Model (SADM) [14] introduced a transformer-conditioned diffusion process to enable autoregressive generation of future medical images based on prior observations. Unlike conventional models relying on a single image for prediction, SADM captures temporal dependencies across sequences, addressing challenges related to missing data, variable sequence lengths, and anatomical consistency.

1.7.2 Simulation of Aging and Disease Progression

Beyond direct imputation, a growing body of research focuses on synthesizing anatomically realistic trajectories of aging and disease progression. These models aim to predict plausible future states of brain morphology, enabling the study of neurodegenerative processes even when longitudinal data is incomplete.

Several methods apply deformation-based counterfactual augmentation strategies. For instance, Equitable Modelling of Brain Imaging by Counterfactual Augmentation with Morphologically Constrained 3D Deep Generative Models [49] proposed generating realistic counterfactual volumetric brain images by applying diffeomorphic deformations conditioned on contextual variables such as age and sex. This technique enables biologically informed data augmentation without explicit temporal supervision.

Latent diffusion models have also been adapted for disease trajectory simulation. The work Enhancing Spatiotemporal Disease Progression Models via Latent Diffusion and Prior Knowledge [50] integrates Latent Diffusion Models (LDMs) with ControlNet-based conditioning, enabling subject-specific morphological synthesis guided by disease priors. By embedding anatomical control signals, the model captures disease progression patterns across brain regions while maintaining anatomical fidelity.

Alternative approaches have tackled the challenge of modeling brain aging without explicit longitudinal data. Learning to Synthesize the Ageing Brain Without Longitudinal Data [51] trained a generative model conditioned on continuous variables such as brain age and disease stage, synthesizing future-aged versions of brain MRIs based solely on cross-sectional inputs. Using stacked 2D slice synthesis to reconstruct full 3D volumes, the method bypasses the need for temporally paired scans while still modeling plausible anatomical evolution.

1.7.3 Challenges and Gaps in existing methodology

Despite notable progress, several challenges persist. Diffusion-based methods, while generating anatomically plausible images, are computationally intensive, with inference times reaching several minutes for a single 3D volume. Furthermore, most current frameworks treat image imputation and disease trajectory simulation separately, rather than unifying them within a single modeling paradigm. Thus, there is a need for developing bidirectional and anatomically constrained generative frameworks capable of simultaneously imputing missing scans and simulating future disease progression in a temporally coherent manner.

1.8 Objective and Scope of the Proposal

The objective of this thesis is to develop anatomically grounded frameworks for imputing missing longitudinal brain MRI scans, modeling individualized trajectories of brain aging and neurodegeneration, and analyzing generated anatomical sequences to better understand disease progression. By reconstructing incomplete imaging sequences in a clinically plausible and anatomically coherent manner, this work aims to improve the ability to track structural brain changes over time, particularly in the context of neurodegenerative diseases such as Alzheimer's disease. Central to this objective is the focus on deformation-based modeling, wherein subject-specific anatomical evolution is captured through physiologically realistic transformations rather than direct intensity-based synthesis. This modeling approach preserves spatial fidelity, ensures biological plausibility, and enables the reconstruction of temporally continuous imaging sequences that can support downstream clinical analysis and disease monitoring.

Beyond imputation, this thesis also emphasizes learning latent representations that disentangle the underlying factors contributing to brain morphological change, such as aging, disease progression, and temporal ordering. By structuring the latent space with explicit conditioning on clinical metadata, such as subject age and diagnostic status, the goal is to enhance the interpretability and predictive robustness of modeled trajectories. This latent structure enables interpolation between observed timepoints and facilitates temporally aware modeling of subject-specific anatomical changes.

In addition to the generative modeling components, this work adopts an "Analysis by Synthesis" paradigm, wherein the generated anatomical data are systematically analyzed to assess biological plausibility, trajectory consistency, and subgroup differences. The evaluation framework includes both image-level assessments—such as anatomical similarity metrics—and trajectory-level analyses, such as the consistency of volume changes across regions of interest (ROIs) and demographic stratifications by age, gender, and disease group. This analytical approach provides deeper insights into the quality and clinical relevance of the

generated longitudinal sequences, moving beyond traditional image reconstruction metrics toward a more comprehensive understanding of modeled brain evolution.

The scope of this thesis encompasses anatomically constrained image synthesis, temporally structured latent modeling, and trajectory-level evaluation in longitudinal brain imaging. All experiments are conducted on structural T1-weighted MRI datasets, with primary evaluations based on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The focus remains on modeling structural changes associated with healthy aging and Alzheimer's disease progression, using subject-specific brain imaging sequences to validate the proposed frameworks. Through anatomically grounded synthesis, temporally disentangled modeling, and detailed trajectory analysis, this thesis aims to contribute new methodologies for understanding individualized patterns of brain aging and neurodegeneration.

1.9 Contributions

This thesis advances the field of longitudinal brain MRI modeling through the following technical contributions:

- Development of a trajectory-conditioned, deformation-based synthesis framework that imputes missing intermediate scans by jointly leveraging past and future anatomical context across populations, enabling more accurate subject-specific longitudinal reconstruction.
- Design of a compact latent space for stationary velocity fields (SVFs), learned from synthetically augmented deformations, which facilitates effective learning under limited data settings by decoupling deformation modeling from raw image intensity generation.
- Introduction of a spatial refinement module that enhances the fidelity of predicted deformation fields, preserving fine-grained anatomical structures in the synthesized brain scans.
- Formulation of a metadata-informed variational autoencoder (VAE) architecture for latent SVF modeling, enabling the disentanglement of temporal, demographic, and disease-related morphological variations within a unified representation space.
- Development of a metadata-aware deformation synthesis strategy through the integration of clinical metadata, such as age and disease status, improving the subject-specific accuracy of synthesized anatomical trajectories.

1.10 Structure of the Proposal

Following the introductory chapter, which outlined the motivation, challenges, and goals of this thesis, the subsequent chapters systematically develop the proposed frameworks and present the experimental evaluations.

Chapter 2 introduces SynBADD, a deformation-based framework designed for longitudinal image imputation. This chapter begins by defining the objective of SynBADD and elaborates on how morphological changes over time can be modeled through piecewise stationary velocity fields (SVFs). It further discusses the rationale behind using SVFs for capturing brain atrophy patterns, particularly in the context of Alzheimer's disease, and details the process of learning subject-specific anatomical trajectories. Subsequent sections present the synthesis of latent deformations through synthetic augmentation and the modeling of individualized temporal trajectories, culminating in an anatomically grounded approach to image generation.

Building on the foundation laid in Chapter 2, Chapter 3 presents DIVA, a metadata-informed variational autoencoder tailored for encoding temporal SVF dynamics. This chapter focuses on enhancing the learning space through synthetic intra-subject augmentation and emphasizes the disentanglement of temporal features within the latent space. It also explores metadata-informed conditioning strategies that incorporate clinical variables such as age and disease status directly into the encoding process. Finally, the chapter introduces the concept of temporal latent arithmetic, enabling interpolation and prediction of subject-specific anatomical progression over time.

Chapter 4 shifts the focus toward an "Analysis by Synthesis" framework, where the quality and clinical relevance of the generated data are systematically evaluated. This chapter analyzes volumetric features extracted from synthesized scans, investigates demographic variations in synthesized brain volumes across age and disease groups, and models anatomical trajectories using the latent SVF representations learned previously. Through these evaluations, Chapter 4 provides insight into the biological plausibility and utility of the proposed models for understanding brain aging and neurodegenerative progression.

Finally, Chapter 5 concludes the thesis by summarizing the key contributions, discussing limitations of the current work, and outlining future directions. It reflects on how anatomically grounded synthesis and trajectory modeling approaches can contribute to advancing longitudinal neuroimaging analysis and enable deeper understanding of individualized patterns of brain aging and disease evolution.

Chapter 2

SynBADD: Longitudinal Medical Image Imputation Via Synthetic Bidirectional Anatomy-guided Diffeomorphic Deformations

2.1 Introduction and Motivation

As outlined in Section 1.1 and Section 1.2, longitudinal brain MRI is critical for understanding the temporal dynamics of neurodegeneration, yet is often hindered by missing scans and inconsistent temporal sampling [6, 7]. Addressing the imputation of missing intermediate scans is essential for enabling accurate modeling of subject-specific anatomical trajectories, particularly in the context of Alzheimer's disease (AD), where early detection of morphological changes is crucial [4, 5].

While Section 1.3 discussed various generative frameworks for brain MRI synthesis, including GANs, VAEs [52], and diffusion models [11, 12], these intensity-based methods often suffer from high computational demands and lack explicit anatomical constraints, limiting their clinical utility. Moreover, as detailed in Section 1.7, existing imputation methods typically predict future scans unidirectionally, failing to fully leverage the rich contextual information present in both past and future timepoints [14].

In response to these challenges, this chapter introduces SynBADD, a deformation-based framework for anatomically grounded longitudinal image imputation. Instead of generating raw voxel intensities, SynBADD models missing scans via stationary velocity fields (SVFs), capturing smooth, anatomically plausible deformations. This deformation-space modeling preserves anatomical topology, mitigates artifacts associated with intensity synthesis, and enables the generation of intermediate scans that are both spatially consistent and anatomically plausible.

A key tool for assessing the anatomical plausibility of deformations is the Jacobian determinant of the transformation field ϕ, denoted as $\det(J_\phi(x))$ at spatial location x. The Jacobian matrix $J_\phi(x)$ at point x is defined as the spatial gradient of the deformation ϕ:

$$J_\phi(x) = \nabla\phi(x) = \begin{bmatrix} \frac{\partial\phi_1}{\partial x_1} & \frac{\partial\phi_1}{\partial x_2} & \frac{\partial\phi_1}{\partial x_3} \\ \frac{\partial\phi_2}{\partial x_1} & \frac{\partial\phi_2}{\partial x_2} & \frac{\partial\phi_2}{\partial x_3} \\ \frac{\partial\phi_3}{\partial x_1} & \frac{\partial\phi_3}{\partial x_2} & \frac{\partial\phi_3}{\partial x_3} \end{bmatrix}.$$

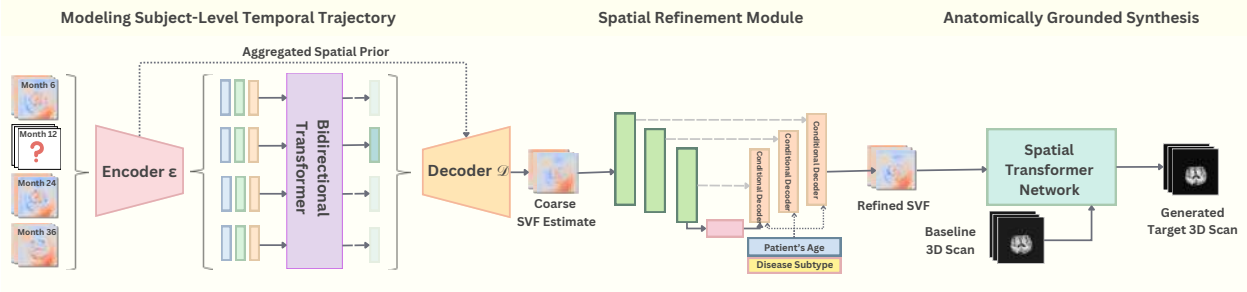


Figure 2.1: An overview of the proposed SynBADD framework, including learning compact SVF representations, modeling subject-specific temporal trajectories, refining coarse SVF estimates, and synthesizing missing scans by warping the baseline image. (Adapted from our work SynBADD)

The determinant of this matrix, $\det(J_\phi(x))$, quantifies the local volume change induced by the deformation at point x.

Interpretations of the Jacobian determinant values are as follows: - $\det(J_\phi(x)) = 1$ indicates local volume preservation (no expansion or shrinkage), - $\det(J_\phi(x)) > 1$ indicates local expansion, - $\det(J_\phi(x)) < 1$ indicates local shrinkage or atrophy.

In the context of modeling brain aging and disease, a realistic deformation field should yield Jacobian determinants close to one in regions of preserved anatomy, while capturing mild localized shrinkage (atrophy) in regions vulnerable to neurodegeneration, such as the hippocampus, parahippocampal gyrus, and third ventricle [4, 5]. Excessive distortion (very high or negative Jacobians) would indicate non-physiological deformations.

Our experiments demonstrate that SynBADD maintains Jacobian distributions within clinically plausible ranges between 0.92 to 0.93 [53], confirming the anatomical soundness of imputed scans. Furthermore, SynBADD achieves substantial improvements in the reconstruction of volumes in neurodegeneration critical regions, particularly the hippocampus, parahippocampal gyrus, and third ventricle, reflecting its ability to model subject-specific disease trajectories with high anatomical fidelity.

Through bidirectional temporal reasoning, deformation-based synthesis, and anatomically constrained regularization, SynBADD addresses key limitations in existing longitudinal modeling approaches. The following sections detail the underlying methodology, including latent deformation learning, temporal trajectory modeling, and anatomically grounded synthesis strategies.

2.2 Methodology

This section details the SynBADD framework for anatomically grounded longitudinal brain MRI imputation. Our approach consists of three primary stages: learning a compact latent space of stationary velocity fields (SVFs) from synthetically augmented deformations, modeling subject-specific temporal trajectories through bidirectional reasoning, and synthesizing anatomically consistent missing scans via refined deformation fields. An overview of the full SynBADD pipeline, including latent deformation learning, temporal trajectory modeling, and scan synthesis, is illustrated in Figure 2.1.

2.2.1 Latent Deformations via Synthetic SVFs

A central component of SynBADD is the learning of a compact latent representation for subject-specific deformations. Rather than modeling intensity values directly, we operate in the deformation space, learning latent embeddings of stationary velocity fields (SVFs) corresponding to anatomical changes relative to a baseline scan.

We employ a 3D U-Net style Variational Autoencoder (VAE) architecture [52] with skip connections. Each registered SVF $\phi_{T_i \to 0}$, computed between a longitudinal scan at time T_i and the baseline scan T_0, is independently encoded into a latent vector $z_i \in \mathbb{R}^d$ along with the corresponding multi-resolution skip features s_i:

$$(z_i, s_i) = \mathcal{E}(\phi_{T_i \to 0}),$$

where $\mathcal{E}$ denotes the encoder. The decoder $\mathcal{D}$ reconstructs the SVF from the latent code and skip features:

$$\hat{\phi}_{T_i \to 0} = \mathcal{D}(z_i, s_i).$$

The VAE is trained using a standard variational loss consisting of a reconstruction loss (mean squared error) and a Kullback-Leibler (KL) divergence penalty to encourage latent distribution regularization.

To overcome the limitations of small longitudinal datasets, we introduce an on-the-fly synthetic augmentation strategy during VAE training. Augmented SVFs are generated through random combinations of:

- Random affine transformations (scaling, rotation, shearing) preserving diffeomorphism,
- Elastic deformations constrained to maintain smoothness,
- Small random noise perturbations.

These augmentations enrich the intra-subject deformation space, allowing the model to generalize better across unseen subjects and anatomical variations while maintaining physiological realism.

During inference, when imputing a missing timepoint T_p, we leverage the available SVF skip features $\{s_j\}_{j \neq p}$ to compute an aggregated spatial prior:

$$\bar{s} = \frac{1}{n-1} \sum_{j \neq p} s_j,$$

which provides a context-aware initialization for decoding the missing deformation from its predicted latent vector $\hat{z}_p$.

By disentangling spatial deformation modeling from direct intensity prediction and enriching the training space with anatomically plausible augmentations, this stage establishes a compact yet expressive latent representation critical for robust temporal modeling.

The VAE-based architecture used for learning compact SVF representations, along with the synthetic augmentation strategy, is shown in Figure 2.2.

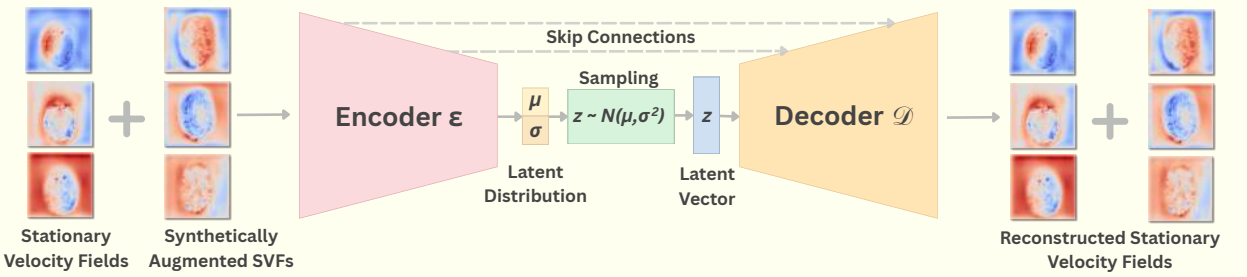


Figure 2.2: Learning compact latent representations of stationary velocity fields (SVFs) with synthetic augmentation for improved deformation modeling. (Adapted from our work SynBADD)

2.2.2 Modeling Subject-Specific Temporal Trajectories

Once compact latent representations of anatomical deformations have been obtained, the next challenge is to model how these deformations evolve over time for each subject. SynBADD addresses this by introducing a bidirectional transformer-based model that captures subject-specific temporal trajectories within the latent deformation space.

Given a set of latent codes $\{z_j\}_{j \neq p}$ corresponding to observed timepoints surrounding the missing scan at timepoint T_p, the objective is to predict the latent code $\hat{z}_p$ representing the deformation from the baseline scan to the missing anatomical state. SynBADD leverages information from both past and future observations relative to the missing timepoint, explicitly conditioning on temporal directionality to enhance prediction accuracy. To this end, each latent code is augmented with learnable time-dependent embeddings that differentiate between past, future, and missing timepoints. This temporal encoding framework enables the model to capture asymmetries in anatomical progression and supports more context-aware and anatomically coherent imputation of missing scans.

To model temporal dependencies, we employ a transformer architecture [54], augmented with domain-specific temporal embeddings. Each input latent code z_j is enriched with:

- **Positional Embeddings (PE):** Encoding the relative ordering of timepoints within a subject's sequence.
- **Time Directionality Embeddings (TDE):** Explicitly distinguishing whether the input timepoint occurs before or after the missing scan, thereby enabling the model to learn different weighting patterns for past and future information.

The transformer encoder processes the sequence of temporally embedded latent codes, producing a context vector that summarizes the subject-specific trajectory. A prediction head then regresses this context vector into the latent code $\hat{z}_p$ of the missing scan:

$$\hat{z}_p = \mathcal{F}(\{z_j + \mathrm{PE}_j + \mathrm{TDE}_j\}_{j \neq p}),$$

where $\mathcal{F}$ denotes the transformer-based trajectory model.

This bidirectional modeling approach allows SynBADD to flexibly interpolate between anatomical states, capturing complex, nonlinear progression patterns that are characteristic of neurodegenerative diseases. By

leveraging both past and future anatomical context, the model achieves more accurate and anatomically coherent imputation compared to traditional unidirectional frameworks.

The overall sequence modeling strategy, including the bidirectional temporal transformer and latent code prediction, is illustrated within the SynBADD architecture diagram in Figure 2.1.

2.2.3 Spatial Refinement via U-Net Module

Following the prediction of the latent code $\hat{z}_p$, SynBADD synthesizes the missing anatomical scan through a two-stage deformation refinement and warping process.

The decoded stationary velocity field (SVF) $\tilde{\phi}_{T_p \to 0}$ is refined using a full-resolution U-Net deformation generator $\mathcal{G}$ [55], dynamically conditioned on subject-specific clinical metadata. Specifically, the subject's age and diagnosis label are injected into each layer of $\mathcal{G}$ using dynamic affine feature-map transform (DAFT) conditioning [56]. This metadata-informed conditioning allows the refinement process to adjust deformation patterns according to clinically relevant attributes as well as normalize and scale the velocities.

The final refined SVF $\hat{\phi}_{T_p \to 0}$ is obtained as:

$$\hat{\phi}_{T_p \to 0} = \mathcal{G}(\tilde{\phi}_{T_p \to 0}, \text{age}, \text{label}).$$

2.2.4 Anatomically Grounded Image Synthesis

The refined deformation field is then used to warp the baseline scan I_0 using a spatial transformer network (STN) [57], producing the synthesized scan $\hat{I}_{T_p}$:

$$\hat{I}_{T_p} = \text{STN}(I_0, \hat{\phi}_{T_p \to 0}).$$

All deformation and synthesis operations are performed at full spatial resolution to preserve subtle anatomical details essential for clinically relevant neuroimaging analysis.

Training of the refinement and synthesis pipeline is supervised by a composite weighted objective, combining both deformation-level and image-level losses:

- **Deformation Energy Loss:** Inspired by LDDMM formulations [58], encouraging smooth, diffeomorphic deformations and penalizing excessive deformation energy (weight = 0.5).
- **Structural Similarity (SSIM) and Mean Squared Error (MSE) Losses:** Promoting anatomical and intensity consistency between synthesized and ground-truth scans (each weighted at 0.05).
- **Jacobian Determinant Regularization:** A regularization term that encourages the determinant of the Jacobian $\det(J_{\hat{\phi}}(x))$ to center around 0.9, reflecting biologically realistic patterns of atrophy commonly observed in neurodegenerative progression [53].

Through anatomically constrained refinement, metadata-informed conditioning, and anatomically motivated regularization, SynBADD achieves anatomically faithful and clinically plausible synthesis of missing longitudinal brain MRI scans.

2.3 Experimental Settings

This section describes the datasets, preprocessing protocols, network architectures, and training procedures employed to evaluate the proposed SynBADD framework for longitudinal brain MRI imputation.

2.3.1 Dataset and Preprocessing

We evaluate SynBADD on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [59], a large-scale, publicly available longitudinal imaging cohort designed for studying Alzheimer's disease progression. We utilize structural T1-weighted MRI scans from subjects spanning cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) diagnostic groups.

Each subject contributes between two and six scans, corresponding to scheduled visits at screening, 6 months, 12 months, 18 months, 24 months, and 36 months. Thus, the longitudinal span covers up to three years of follow-up, with approximately six-month intervals between acquisitions.

All scans were preprocessed using FreeSurfer v8.0.0 [60]. Preprocessing steps included:

- **Bias Field Correction:** B1 bias field correction followed by N3 non-uniformity correction were applied to reduce intensity inhomogeneities.
- **Intensity Normalization:** Intensity scaling was performed using `mri_convert`. If scaling metadata was missing, a uniform scaling factor of 1.0 was used.
- **Skull Stripping:** Skull stripping was applied using `mri_watershed` for scans that were not already skull-stripped.
- **Rigid Registration:** Each scan was rigidly aligned to the corresponding subject's baseline scan using ANTs, restricting transformations to translation and rotation to preserve individual brain morphology.
- **Volume Resampling:** All scans were zero-padded to a uniform cubic volume of size $(256\times256\times256)$ and subsequently downsampled to $(128 \times 128 \times 128)$ to maintain anatomical fidelity while reducing computational cost.
- **Deformation Estimation:** Stationary velocity fields (SVFs) representing the deformations from the baseline to each follow-up scan were computed using VoxelMorph [58], an unsupervised diffeomorphic registration framework.

Finally, all scans were intensity normalized to have zero mean and unit variance. This comprehensive preprocessing pipeline ensures consistent spatial alignment across timepoints while preserving subject-specific anatomical features, allowing SynBADD to model true biological changes rather than technical variability.

Parcellations: To ensure the accurate delineation of each region of interest (ROI), parcellation was performed on the original-resolution images, prior to skull-stripping and the application of any pre-processing steps. Subsequently, the resulting parcellation maps were subjected to the same pre-processing pipeline as

the MRI volumes to maintain consistency across data representations. Finally, the parcellations were registered to the corresponding MRI images using nearest-neighbour interpolation, thereby preserving discrete label integrity and ensuring correct spatial correspondence between the parcellations and volumetric data.

2.3.2 Implementation Details

2.3.2.1 Network Architectures

The VAE encoder and decoder architectures are based on 3D U-Net backbones [55], with five convolutional layers and multi-resolution skip connections. During VAE pretraining, we apply synthetic deformation augmentations designed to preserve diffeomorphism, including random 3D rotations from the SO(3) group (based on Euler angles), affine ramp fields, and elastic noise scaling.

The temporal trajectory model is a transformer-based encoder [54] with two layers and eight attention heads, operating over a sequence length of five latent codes corresponding to longitudinal scans following the baseline. Each latent code is enriched with positional embeddings and time directionality embeddings before being processed.

The deformation refinement network is a 3D U-Net conditioned at every layer via dynamic affine feature-map transforms (DAFT) [56] based on subject age and diagnosis metadata. Spatial warping of the baseline scan is performed using a differentiable spatial transformer network (STN) [57].

2.3.2.2 Training Procedure

We train all models using four NVIDIA RTX 6000 GPUs (48 GB memory each). A five-fold cross-validation strategy is employed, ensuring that each subject appears in exactly one of the train, validation, or test sets to prevent data leakage. In each fold, approximately 1100 training samples, 120 validation samples, and 320 test samples are included. This corresponds to 275 subjects for train split, 30 subjects for validation split and 77 subjects for test split where each subject has between 3 and 5 timepoints available.

The optimizer is Adam with an initial learning rate of 5×10^{-5}, coupled with a Reduce-on-Plateau scheduler (patience = 5 epochs) and early stopping (patience = 7 epochs) based on validation loss. The batch size is set to 6 for all experiments.

During VAE pretraining, we apply on-the-fly synthetic augmentations, including:

- Random 3D SO(3) rotations,
- Smooth affine ramp fields,
- Elastic noise perturbations.

These augmentations enrich the intra-subject deformation space while maintaining anatomical realism.

The best-performing model checkpoint is selected based on the lowest validation loss during training, and all final evaluations are reported on the held-out test set.

Variants	Design Components				Metrics		
	PE	TDE	Ref.	Clin.	PSNR↑	SSIM↑	VXM↓
V0	✗	✗	✓	✗	32.7588	0.981764	77.1520
V1	✓	✗	✓	✗	32.7326	0.981742	77.1260
V2	✓	✓	✓	✗	32.7650	0.981822	76.9560
V3	✓	✓	✗	✗	32.0353	0.979381	78.1640
V4	✓	✓	✓	✓	**32.8545**	**0.981943**	**76.7000**

Table 2.1: Impact of design choices on missing scan imputation. PE: positional embedding, TDE: time directionality embedding, Ref.: refinement, Clin.: clinical conditioning. Metrics include PSNR, SSIM, and VoxelMorph diffeomorphic registration loss (VXM).

2.4 Results and Analysis

This section presents the experimental results evaluating SynBADD's performance for missing longitudinal scan imputation. We conduct detailed ablation studies to assess the impact of key design components, benchmark SynBADD against state-of-the-art baselines, and provide qualitative evaluations highlighting anatomical plausibility and temporal consistency.

2.4.1 Ablation Studies

To evaluate the contribution of individual architectural components, we perform ablation studies focusing on the following design factors:

- **Positional Embedding (PE):** Encoding temporal order information across observed scans.
- **Time Directionality Embedding (TDE):** Differentiating past versus future observations relative to the missing timepoint.
- **Refinement Module:** Spatial refinement of the decoded SVF using a U-Net architecture conditioned on clinical metadata.
- **Clinical Metadata Conditioning (using DAFT):** Integrating demographic and diagnostic attributes into the deformation refinement process.

We systematically disable subsets of these components to create five model variants (V0–V4) and evaluate their impact on missing scan imputation. Performance is assessed using three quantitative metrics: - Peak Signal-to-Noise Ratio (PSNR), - Structural Similarity Index Measure (SSIM), - VoxelMorph registration loss (VXM).

Results are summarized in Table 2.1.

From Table 2.1, we observe the following: - Introducing positional embeddings (PE) alone does not significantly affect PSNR or SSIM, but slightly improves registration consistency (lower VXM loss). - Incorporating time directionality embeddings (TDE) alongside PE further improves deformation quality, reducing VXM loss and enhancing anatomical realism. - Removing the refinement module results in a

Metric	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD
Mean PSNR	29.0959	**32.2085**	29.9985	**33.2227**	30.0095	**33.2323**	29.4015	**32.5388**	29.9715	**33.0703**
Mean SSIM	0.9510	**0.9778**	0.9612	**0.9840**	0.9621	**0.9836**	0.9586	**0.9803**	0.9623	**0.9840**

Table 2.2: Mean PSNR and SSIM values across five folds for missing scan imputation benchmarking experiments. Higher scores are highlighted in bold.

notable drop in performance across all metrics, confirming the importance of spatial refinement for preserving fine-grained anatomy. - Clinical metadata conditioning (Clin.) provides an additional improvement in both PSNR and SSIM, suggesting that subject-specific factors such as age and disease status help guide anatomically plausible deformation prediction.

Overall, the full model (V4) incorporating all design choices achieves the best performance across all evaluation metrics, validating the effectiveness of SynBADD's modular architecture.

2.4.2 Benchmark Comparisons

We benchmark SynBADD against two state-of-the-art baselines for missing longitudinal brain MRI scan imputation:

- **ReMiND** [15]: A diffusion-based model that generates missing scans slice-by-slice and stacks them to reconstruct 3D volumes.
- **NODER** [18]: A subject-wise neural ODE framework that models temporal anatomical trajectories through deformation-based interpolation which has to be trained separately for each target timepoint.

Performance is evaluated quantitatively using three primary metrics: - **Peak Signal-to-Noise Ratio (PSNR)**, - **Structural Similarity Index Measure (SSIM)**, - **Dice similarity coefficient** between anatomical segmentations warped back to baseline.

In addition, we analyze the Jacobian determinant statistics of the generated deformations to assess anatomical plausibility.

2.4.2.1 Quantitative Performance

Table 2.2 reports the mean PSNR and SSIM values across five cross-validation folds. SynBADD consistently outperforms NODER across all folds, achieving a mean PSNR improvement of 10.64% dB and a mean SSIM gain of 2.39%. This indicates superior image-level fidelity and structural preservation.

2.4.2.2 Region-wise Anatomical Fidelity

Table 2.3 summarizes Dice scores for critical neurodegeneration-associated regions: hippocampus, parahippocampal gyrus, and third ventricle. SynBADD significantly outperforms NODER across all regions and folds, achieving notably higher overlap with ground truth segmentations. Improvements in hippocampal and parahippocampal fidelity are especially important given their relevance to Alzheimer's disease progression.

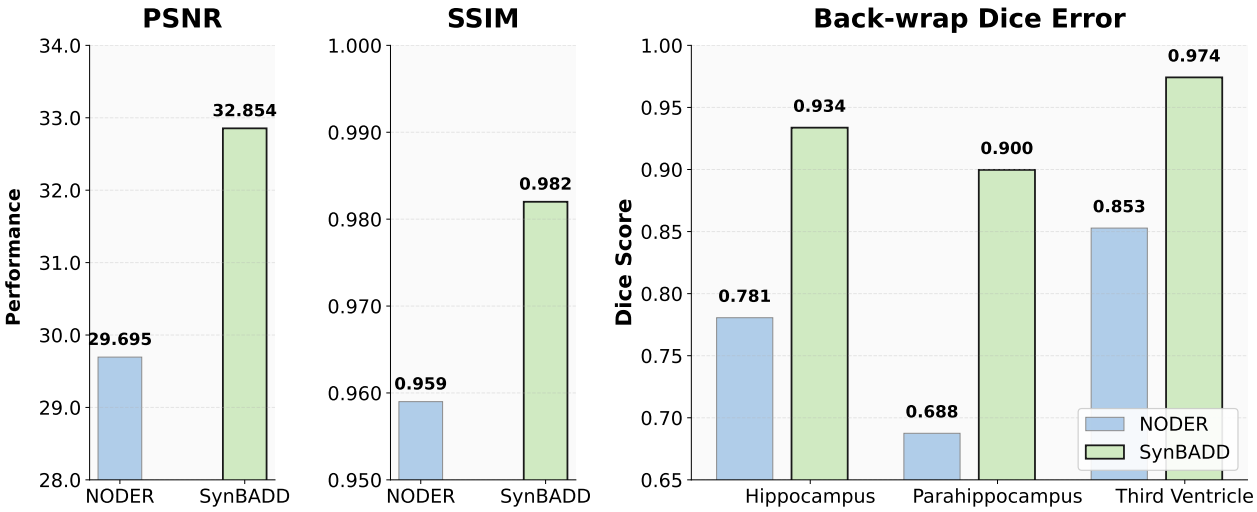


Figure 2.3: Performance comparison against state-of-the-art models on PSNR, SSIM, and back-warp Dice scores across key ROIs.

Region of Interest	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD	NODER	SynBADD
Hippocampus	0.7460	**0.9037**	0.7867	**0.9464**	0.7916	**0.9403**	0.7781	**0.9269**	0.8004	**0.9511**
Parahippocampal	0.6523	**0.8544**	0.7005	**0.9164**	0.7004	**0.9119**	0.6764	**0.8949**	0.7081	**0.9207**
Third Ventricle	0.8483	**0.9634**	0.8555	**0.9851**	0.8545	**0.9716**	0.8471	**0.9683**	0.8584	**0.9828**

Table 2.3: Mean Dice scores for neurodegeneration relevant regions across five folds for missing image imputation. Higher scores are highlighted in bold.

2.4.2.3 Anatomical Plausibility Assessment

Table 2.4 reports the mean and standard deviation of the Jacobian determinant across generated deformations. SynBADD consistently maintains mean Jacobian values close to 0.92–0.93, indicating mild atrophy patterns consistent with neurodegenerative progression while avoiding anatomically implausible expansions or foldings.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Mean Jacobian	0.9309	0.9287	0.9243	0.9292	0.9232
Std Jacobian	0.0400	0.0387	0.0382	0.0399	0.0379

Table 2.4: Mean and standard deviation of Jacobian determinants for SynBADD across test folds.

A summary of overall benchmarking results comparing SynBADD, NODER, and ReMiND across key metrics is shown in Figure 2.3.

2.4.3 Qualitative Evaluation

Beyond quantitative metrics, qualitative assessments provide critical insights into the anatomical plausibility and spatial consistency of the generated scans. In this section, we visually compare the outputs of SynBADD against NODER and ReMiND across multiple subjects.

2.4.3.1 Slice-wise Anatomical Consistency

Figures 2.4 and 2.5 show consecutive axial slices of generated scans for two representative subjects.

ReMiND, a diffusion-based slice-wise generator, exhibits noticeable inconsistencies between adjacent slices, including intensity discontinuities and hole-like artifacts. These issues are particularly prominent between slices 63 and 64, where both image quality and anatomical coherence degrade sharply. Such slice discontinuities compromise the volumetric integrity of reconstructed 3D scans, undermining their clinical reliability.

In contrast, SynBADD, along with NODER, produces spatially consistent volumetric structures across slices. SynBADD's deformation-based synthesis approach ensures that anatomical structures remain continuous across neighboring slices, preserving coherent morphology in critical brain regions.

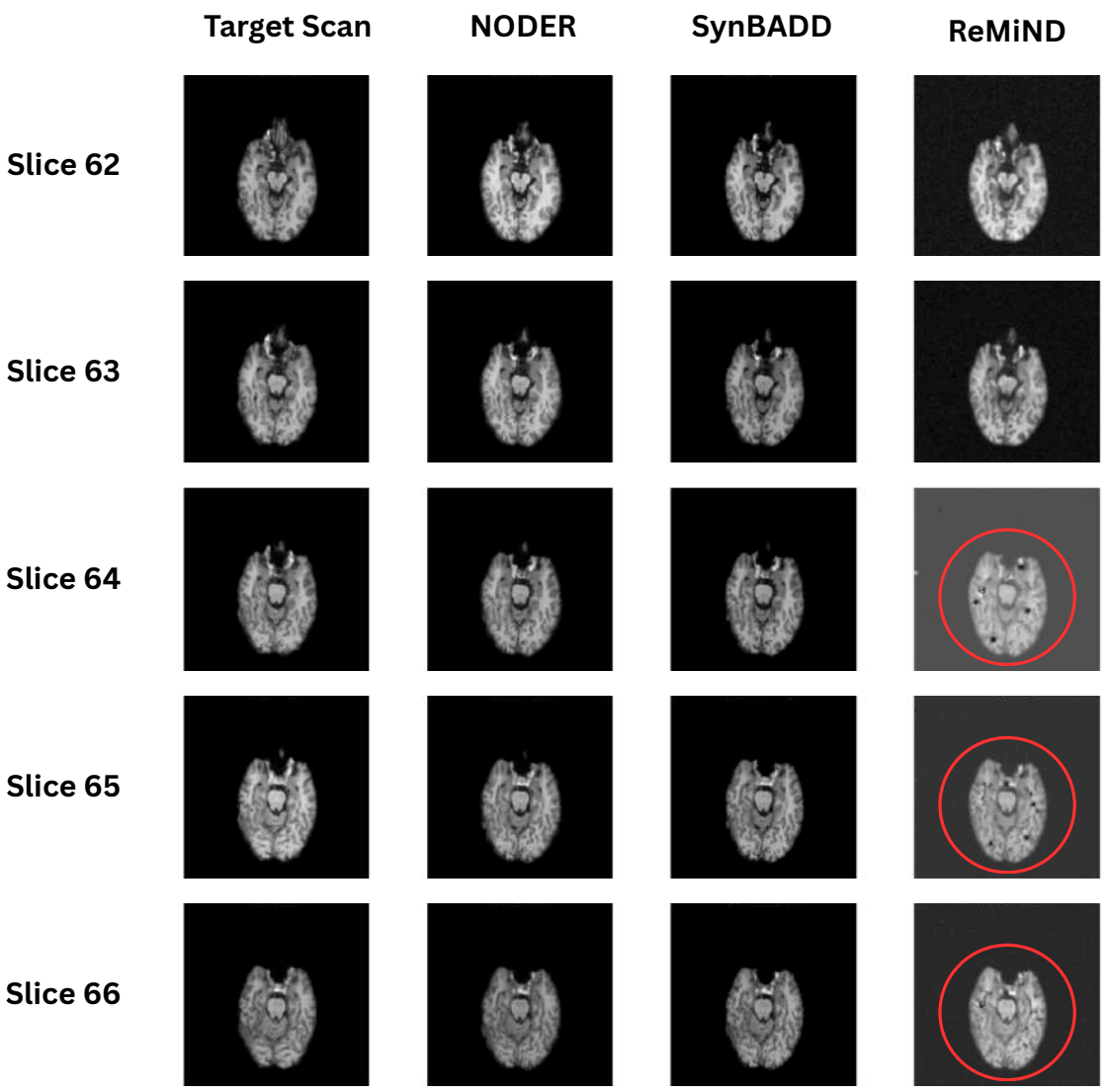


Figure 2.4: 3D spatial consistency analysis across consecutive slices for a subject. ReMiND (diffusion-based, slice-wise generation) exhibits noticeable inconsistencies, with visible intensity differences between consecutive slices and random hole-like artifacts appearing across adjacent slices. These discontinuities are especially prominent between slices 63 and 64, where both signal quality and anatomical coherence are disrupted. In contrast, SynBADD and NODER (deformation-based methods) maintain consistent intensities and preserve continuous 3D anatomical structure across slices.

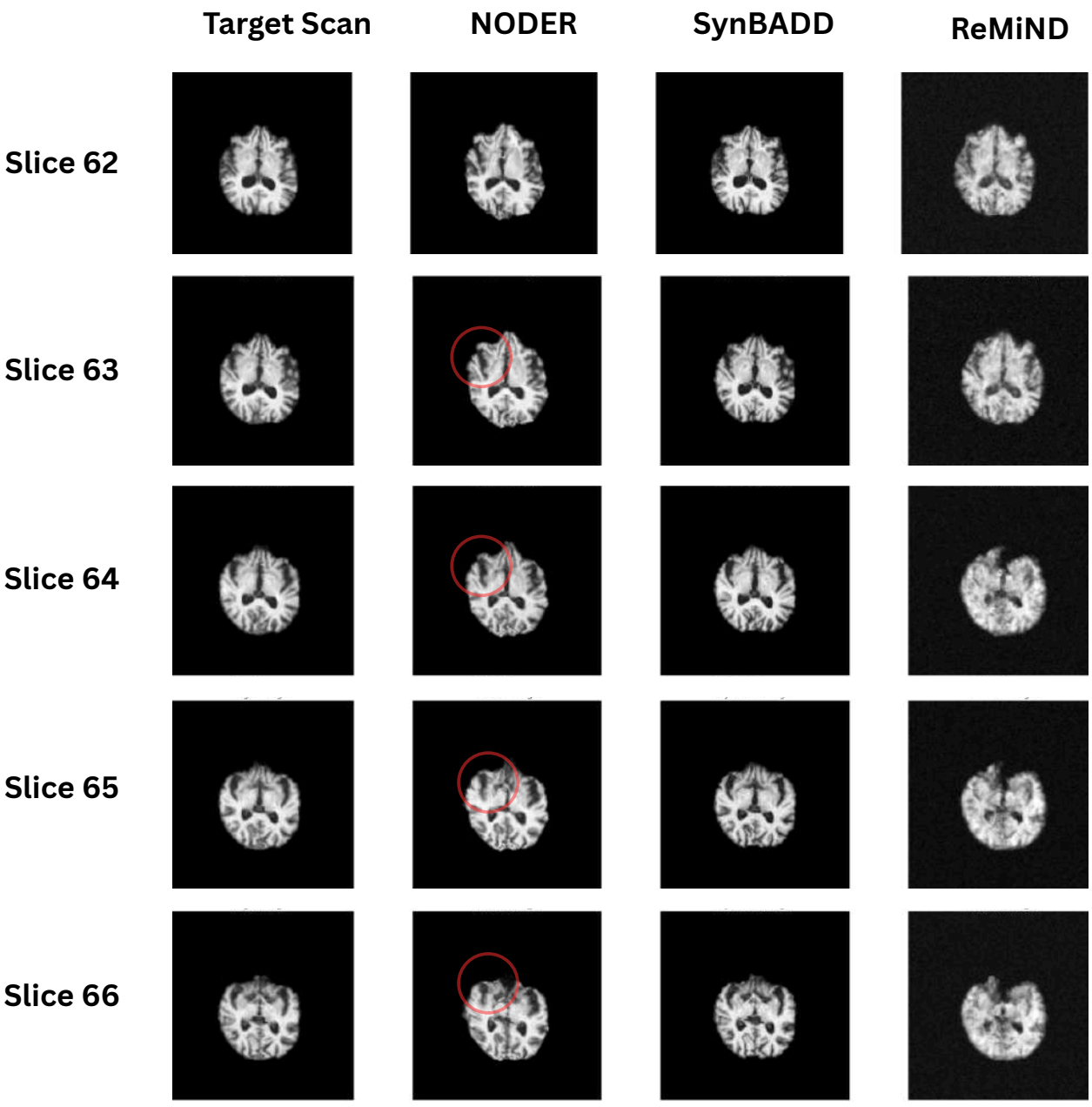


Figure 2.5: Slice-wise generation comparison of SynBADD against baseline models for another subject. SynBADD maintains a coherent 3D anatomical structure across consecutive slices. ReMiND (diffusion-based, slice-wise method) shows loss of spatial coherence between slices 63 and 64, with visible discontinuities. NODER (subject-wise, timepoint-wise deformation model) overestimates local deformations, resulting in a protrusion artifact (highlighted with a red circle).

2.4.3.2 Difference Map Analysis

To further visualize reconstruction fidelity, Figure 2.6 presents voxel-wise mean absolute error (MAE) difference maps between generated and ground-truth scans for each method.

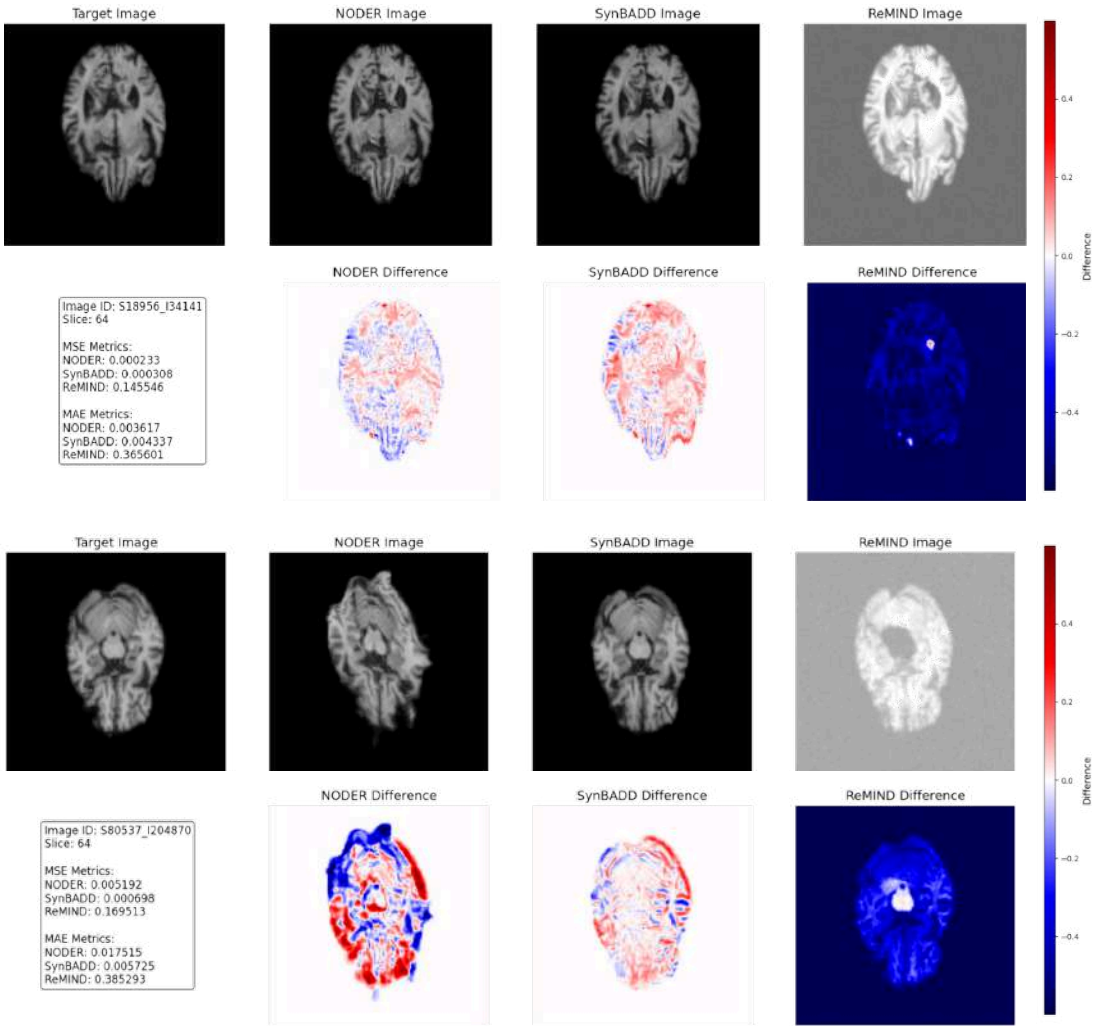


Figure 2.6: Comparison of generated missing scans across different methods. Difference maps show the voxel-wise mean absolute error (MAE) between predicted scans and ground-truth targets. Blue regions indicate negative intensity differences, red regions indicate positive differences, and white corresponds to zero difference. SynBADD achieves significantly lower MAE compared to baseline models. ReMiND (diffusion-based method) introduces artifacts that compromise anatomical fidelity when generating from scratch. NODER tends to overestimate deformations, resulting in exaggerated anatomical changes.

ReMiND produces widespread artifacts, particularly in cortical and ventricular regions, reflecting its challenges in maintaining consistent volumetric structures during slice-wise synthesis. NODER, while generally better, tends to overestimate local deformations, resulting in exaggerated anatomical variations. SynBADD achieves significantly lower MAE values, especially around clinically critical structures such as the hippocampus and ventricular system, indicating more accurate modeling of true anatomical evolution.

These qualitative evaluations reinforce the quantitative findings, demonstrating that SynBADD produces anatomically faithful and spatially coherent synthesized scans, outperforming both diffusion-based and deformation-based baseline methods.

2.5 Discussion

The results presented in the previous sections demonstrate that SynBADD achieves substantial improvements over state-of-the-art baselines in missing scan imputation, both quantitatively and qualitatively. In this section, we discuss the key reasons for SynBADD's superior performance, the advantages of deformation-based synthesis frameworks, and important limitations that highlight opportunities for future work.

2.5.1 Interpretation of Results

SynBADD consistently improves PSNR, SSIM, and Dice scores across folds and regions of interest. The Jacobian determinant distributions further confirm that the deformations generated by SynBADD maintain anatomically plausible transformations, avoiding unrealistic expansions or folding artifacts. Qualitative visualizations reinforce these findings, showing that SynBADD maintains 3D anatomical coherence across slices without introducing discontinuities or deformation exaggerations, unlike slice-wise diffusion models and ODE-based deformation interpolations.

2.5.2 Effectiveness of SVF Latent Modeling

The superior performance of SynBADD is largely attributable to its latent modeling of stationary velocity fields (SVFs). Operating in the deformation space, rather than directly predicting voxel intensities, allows SynBADD to:

- Preserve anatomical topology, avoiding distortions common in intensity-based generative models.
- Capture smooth, anatomically grounded deformation trajectories that align with known neurodegenerative patterns.
- Enable compact, disentangled latent representations that support bidirectional temporal modeling, facilitating anatomically plausible interpolation and imputation.

By decoupling intensity synthesis from deformation modeling, SynBADD leverages the inherent smoothness and structural continuity of anatomical changes, leading to greater robustness across demographic and clinical subgroups.

2.5.3 Advantages of Deformation-based Synthesis

SynBADD offers several important advantages for longitudinal brain MRI imputation. First, it requires only a single training procedure to generalize across diverse subjects and clinical subgroups, significantly reducing computational overhead compared to models that require subject-specific adaptation. Its design

enables robust handling of missing data at inference time, leveraging available observations without retraining. By modeling deformation trajectories within a population-level anatomical space, SynBADD effectively captures subject-specific variations while remaining grounded in the broader statistical structure of healthy and pathological aging.

Unlike continuous regression frameworks, which often suffer from trajectory drift and interpolation artifacts when estimating intermediate anatomical states, SynBADD produces anatomically faithful, temporally discrete predictions aligned with clinically meaningful intervals. Its discretized temporal modeling strategy ensures that imputation focuses specifically on key timepoints, improving both spatial accuracy and clinical interpretability.

Operating in the deformation space confers additional benefits over intensity-based synthesis approaches, such as GANs and diffusion models. Deformation-based methods are computationally more efficient, as generating deformation fields and warping existing scans is substantially less demanding than synthesizing entire 3D volumes voxel-by-voxel. They also inherently preserve spatial coherence across slices, avoiding discontinuities that often arise when generating volumes slice-by-slice. Furthermore, deformation fields can be explicitly regularized using Jacobian determinant constraints, ensuring anatomically plausible transformations that reflect known patterns of brain atrophy and neurodegenerative progression.

Together, these properties make SynBADD and deformation-based synthesis frameworks particularly well-suited for imputation tasks where anatomical fidelity, temporal coherence, and computational scalability are critical requirements.

2.5.4 Limitations and Future Directions

Despite its strengths, SynBADD is subject to several limitations. By design, SynBADD discretizes longitudinal sequences by modeling a fixed number of timepoints relative to a baseline scan. While this ensures stable and anatomically plausible synthesis at specific clinical intervals, it limits the model's ability to capture continuous anatomical evolution or sample arbitrary intermediate timepoints. In contrast, continuous frameworks such as NODER offer theoretical flexibility for timepoint sampling, but often suffer from increased interpolation errors, inaccurate trajectory estimation, and compromised image quality due to their optimization for image regression rather than high-fidelity generation.

Another limitation of SynBADD stems from its reliance on deformation-based modeling, which inherently favors the representation of subtle to moderate anatomical changes. While this approach effectively captures typical neurodegenerative progression, it may struggle to accurately model extreme pathological transformations or the emergence of entirely new anatomical structures, such as tumors, over time. Furthermore, although conditioning on clinical metadata facilitates subject-specific adaptation and enhances the plausibility of synthesized trajectories, it assumes the availability and accuracy of demographic and diagnostic information—an assumption that may not consistently hold in real-world clinical environments due to privacy concerns.

Future work could extend SynBADD to learn time-continuous deformation trajectories over longer longitudinal intervals, such as those available in the full ADNI-2 and ADNI-3 datasets. This would enable finer modeling of subject-specific anatomical changes across broader disease stages. Furthermore, incorporating

hybrid approaches that combine deformation-based warping with targeted local intensity generation could allow the framework to better model the appearance of new tissue structures or extreme volumetric changes. Finally, integrating uncertainty quantification into the generation pipeline could provide clinically valuable confidence estimates for imputed scans, particularly in sparsely sampled or ambiguous temporal regions.

2.6 Summary

This chapter presented SynBADD, a deformation-based framework for missing longitudinal brain MRI imputation that models subject-specific anatomical changes using stationary velocity fields (SVFs) and metadata-informed conditioning. By synthesizing anatomically plausible deformation trajectories rather than generating intensities directly, SynBADD ensures anatomical fidelity, temporal coherence, and robustness to missing data, even in limited data settings. Extensive evaluations demonstrated that SynBADD substantially improves reconstruction quality compared to state-of-the-art baselines, achieving higher fidelity in critical neurodegeneration-relevant regions while preserving biologically realistic atrophy patterns.

Building on these foundations, the next chapter focuses on enhancing the latent deformation modeling process by explicitly disentangling underlying sources of anatomical variation, such as age, disease progression, and temporal evolution. Toward this goal, we introduce DIVA, a metadata-informed variational autoencoder framework designed to enable temporally structured and clinically interpretable latent representations for longitudinal brain MRI analysis.

Chapter 3

DIVA: Disentangled, Metadata-Informed Variational Autoencoder for Temporal SVF Encoding

3.1 Introduction

The previous chapter introduced SynBADD, which demonstrated that deformation-based generative modeling can produce anatomically coherent and temporally plausible brain MRI scans. However, the age and disease label, which influence the trajectory of structural brain changes and help distinguish pathological patterns from normal variation, were incorporated at the refinement stage rather than in the embedding space. As a result, factors such as age-related changes, disease progression, and temporal evolution are encoded without explicit supervision. This may limit the model's ability to produce interpretable representations or generalize across subjects with varying clinical and demographic characteristics. Moreover, temporal entanglement impairs the model's ability to capture consistent anatomical progression, as latent trajectories may conflate temporally ordered changes with static inter-subject variability. Prior work has shown that such entanglement leads to non-smooth or geometrically distorted trajectories in latent space, which degrades the performance of temporal models and hinders disease staging or progression analysis [61–63]. In this chapter, we build on that foundation by focusing on how to structure the latent space of such models to better capture clinically meaningful and temporally ordered variation.

Recent work in generative modeling and clinical neuroimaging highlights the importance of integrating domain knowledge into the latent space of deep models. Disentangled representations have been shown to enhance both interpretability and generalization [64], especially when the underlying variation arises from multiple semantically meaningful sources. In neurodegenerative disease modeling, structural brain changes vary considerably across disease subtypes, progression rates, and ages of onset [63, 65]. Capturing this variation in a structured, disentangled manner is essential not only for accurate image synthesis but also for understanding disease dynamics and staging. [66]

To address these challenges, we introduce **DIVA**—a *Disentangled, metadata-Informed Variational Autoencoder* designed for encoding temporally structured stationary velocity fields (SVFs). DIVA explicitly conditions its latent space on clinically relevant metadata such as age, diagnosis label, and timepoint, encouraging separation between anatomical identity and temporal evolution. This conditioning guides the model to produce more interpretable latent embeddings that better support smooth, consistent, and clinically meaningful trajectory modeling.

Additionally, clinical neuroimaging datasets such as ADNI are often sparse and irregularly sampled across subjects and timepoints. By incorporating metadata into the encoding process, DIVA aligns intra-subject trajectories and enables population-level comparisons, supporting analyses like subgroup stratification across cognitively normal (CN), mildly cognitively impaired (MCI), and Alzheimer's disease (AD) populations [67].

In this chapter, we detail the design and evaluation of DIVA. We show that metadata conditioning improves latent separability across clinical groups, facilitates temporally consistent modeling, and enhances generalization in low-data settings. Through a series of controlled experiments, we analyze how synthetic augmentations, timepoint supervision, and factorized latent design contribute to robust and interpretable SVF encoding.

3.2 Motivation and Experiment Design

Generative models trained on longitudinal neuroimaging data often suffer from entangled representations that conflate spatial, temporal, and subject-specific sources of variation, including clinical metadata such as age or diagnosis. This entanglement limits their usefulness for downstream tasks such as disease staging, scan interpolation, or individualized trajectory modeling. To address these challenges, we experiment with four targeted variations of the DIVA framework, each designed to promote better disentanglement, temporal structure, and integration of relevant clinical information:

1. **Augment intra-subject variation:** Anatomical deformation fields (SVFs) are often limited in number per subject. We introduce high-resolution synthetic augmentations to enrich intra-subject deformation diversity and encourage spatially coherent latent encoding.

2. **Condition on timepoints:** To encourage temporal structure in the latent space, we incorporate timepoint supervision during training. This allows the model to better capture progressive anatomical changes over time and improves alignment across longitudinal scans.

3. **Factorize latent variables:** Simply conditioning on time does not encourage structuring the latent space. Inspired by disentangled representation learning, we explicitly separate each SVF's latent code into a content component and a temporal component. We apply independence and contrastive losses to encourage disentanglement and chronological organization.

4. **Incorporate clinical metadata:** Beyond timepoints, we introduce metadata such as age and diagnosis during training. Conditioning on this information helps disentangle anatomical variation arising from clinical status or demographics, supporting interpretability and group-level analysis.

These variations form the basis for a modular and interpretable VAE architecture. In the following sections, we incrementally build on a vanilla VAE baseline—introducing factorization, contrastive losses, and metadata conditioning—to evaluate how each architectural component affects latent space quality and clinical relevance.

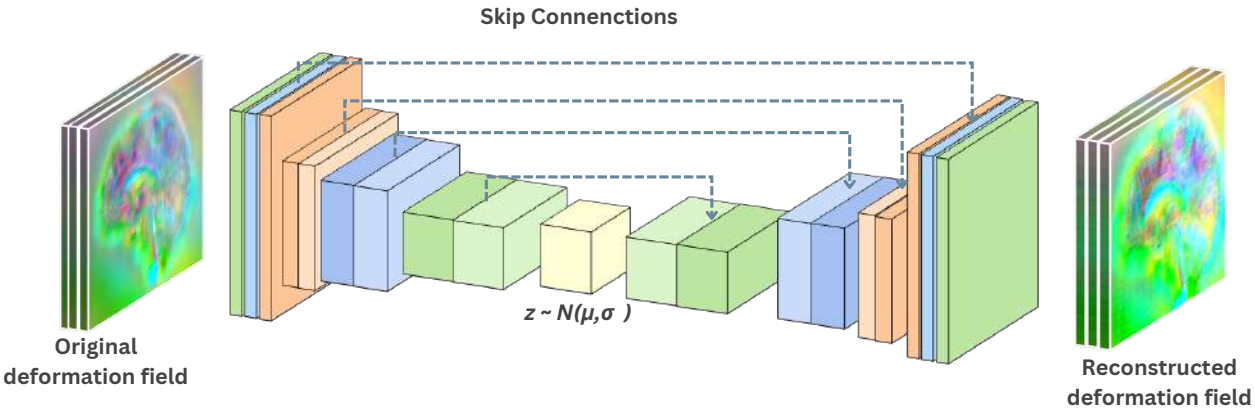


Figure 3.1: Architechture of vanilla 3D VAE with skip connection

3.3 Synthetic Augmentation for Intra-Subject Learning

Longitudinal neuroimaging data is often sparse, with irregular scan intervals and limited anatomical variability per subject. To address this, we first explore how synthetic augmentation alone—without metadata conditioning—affects the learned latent space in a VAE trained to reconstruct stationary velocity fields (SVFs).

We compare two augmentation strategies:

3.3.1 Vanilla VAE with Random Augmentation

The Vanilla-VAE is implemented using a U-Net-style architecture and is designed to reconstruct a stationary velocity field (SVF) from its encoded representation. Each input SVF $\mathbf{x} \in \mathbb{R}^{128\times128\times128\times3}$ is passed through an encoder that parameterizes the approximate posterior distribution $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$, where $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{512}$. During training, a latent variable $\mathbf{z}$ is sampled via the reparameterization trick: $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At inference time, we use $\boldsymbol{\mu}$ as a deterministic embedding. The decoder receives a reshaped projection of $\mathbf{z}$ and progressively upsamples it while integrating skip connections from each corresponding encoder layer. The final output is reconstructed using a $\tanh$ activation to match the SVF intensity range. Refer to Figure 3.1 for an overview of the architechture.

This model is trained on 1,140 real SVFs and 1,000 synthetic samples, where augmentations are generated by applying random 3D rotations, affine ramps, and elastic noise to SVFs drawn uniformly at random from the dataset. These augmentations are agnostic to subject identity and timepoint and thus do not preserve longitudinal structure. In Table 3.1, we report the EPE and angular error along with timepoint classification metrics where the latent embedding obtained from the VAE is passed to a linear classifier to predict which timepoint it belongs to. We report classification accuracy and F1 score for the same. We also perform a pairwise ordering task where given two timepoints we order them chronologically. As seen in the results, although the model achieves low reconstruction loss, the resulting latent space lacks temporal structure and subject-level consistency.

Metric	Value
Test Loss	0.2882
Average End-Point Error (EPE)	0.0584
Average Angular Error (degrees)	8.07
Timepoint Pairwise Ordering Accuracy	0.5113
Timepoint Classification Accuracy	0.1619
Timepoint F1 Score (Macro)	0.1396
Timepoint F1 Score (Weighted)	0.1577
Kendall's τ	0.5124
Spearman's ρ	0.5675

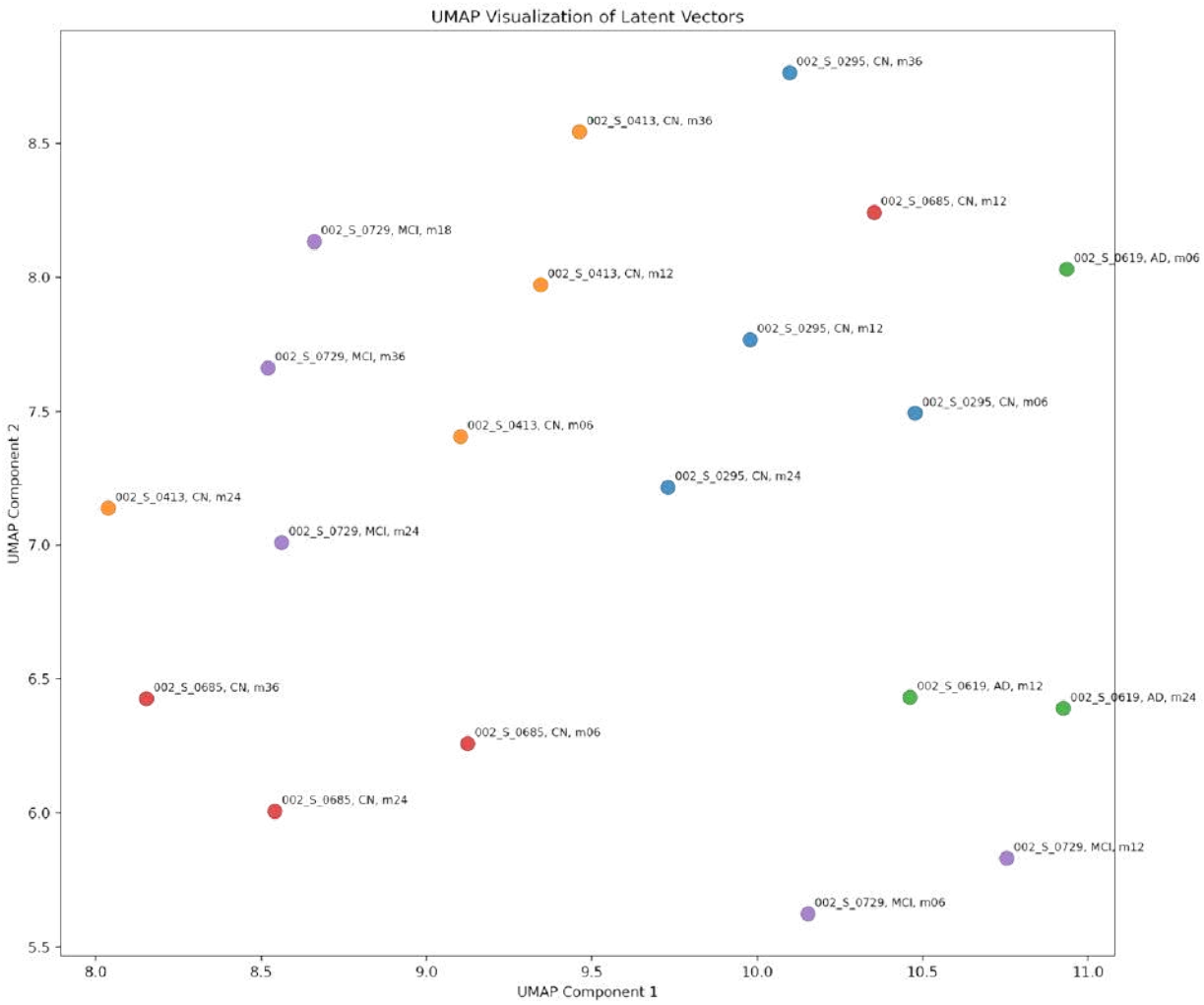
Table 3.1: Evaluation results for the Vanilla VAE model trained with 1k synthetic samples. The latent space shows limited structure and modest classification and ordering performance.

3.3.2 Dense Augmented VAE with Subject-Timepoint Structure

In contrast, the second variant performs structured augmentation by generating 10 synthetic deformations per subject per timepoint, resulting in a training set of 16,900 total samples. This augmentation strategy preserves subject identity and longitudinal context, encouraging the model to capture intra-subject variation more faithfully. While reconstruction quality remains comparable to the baseline (Table 3.2), this model achieves improved directional alignment and subject-level consistency in latent space. A comparision of the latent space between Vanilla and Dense Augmented VAE is shown in Figures 3.2 and 3.3

Metric	Value
Test Loss	0.2891
Average End-Point Error (EPE)	0.0639
Average Angular Error (degrees)	7.25
Timepoint Pairwise Ordering Accuracy	0.5315
Timepoint Classification Accuracy	0.1683
Timepoint F1 Score (Macro)	0.1511
Timepoint F1 Score (Weighted)	0.1700
Kendall's τ	0.5942
Spearman's ρ	0.6182

Table 3.2: Evaluation results for the Dense Augmented VAE model trained with 16k synthetic samples. The model shows improved temporal structure and ordering performance compared to the Vanilla VAE.



Vanilla VAE (1k Synthetic Samples)

Figure 3.2: Visualization of latent space structure in Vanilla VAE model. The Vanilla variant is trained with limited synthetic samples. Each color represents a unique subject.

While structured augmentation enhances reconstruction quality and intra-subject consistency (refer Table 3.13), it alone is insufficient to impose meaningful temporal or clinical structure in the latent space. These limitations motivate the introduction of temporal and metadata-aware conditioning, explored in the next section.

3.4 Latent Conditioning with Timepoint Embeddings for Temporal Structure

Although synthetic augmentation improves spatial and subject-level consistency, it does not introduce temporal ordering or clinical structure in the latent space. To overcome this, we incorporate metadata-aware

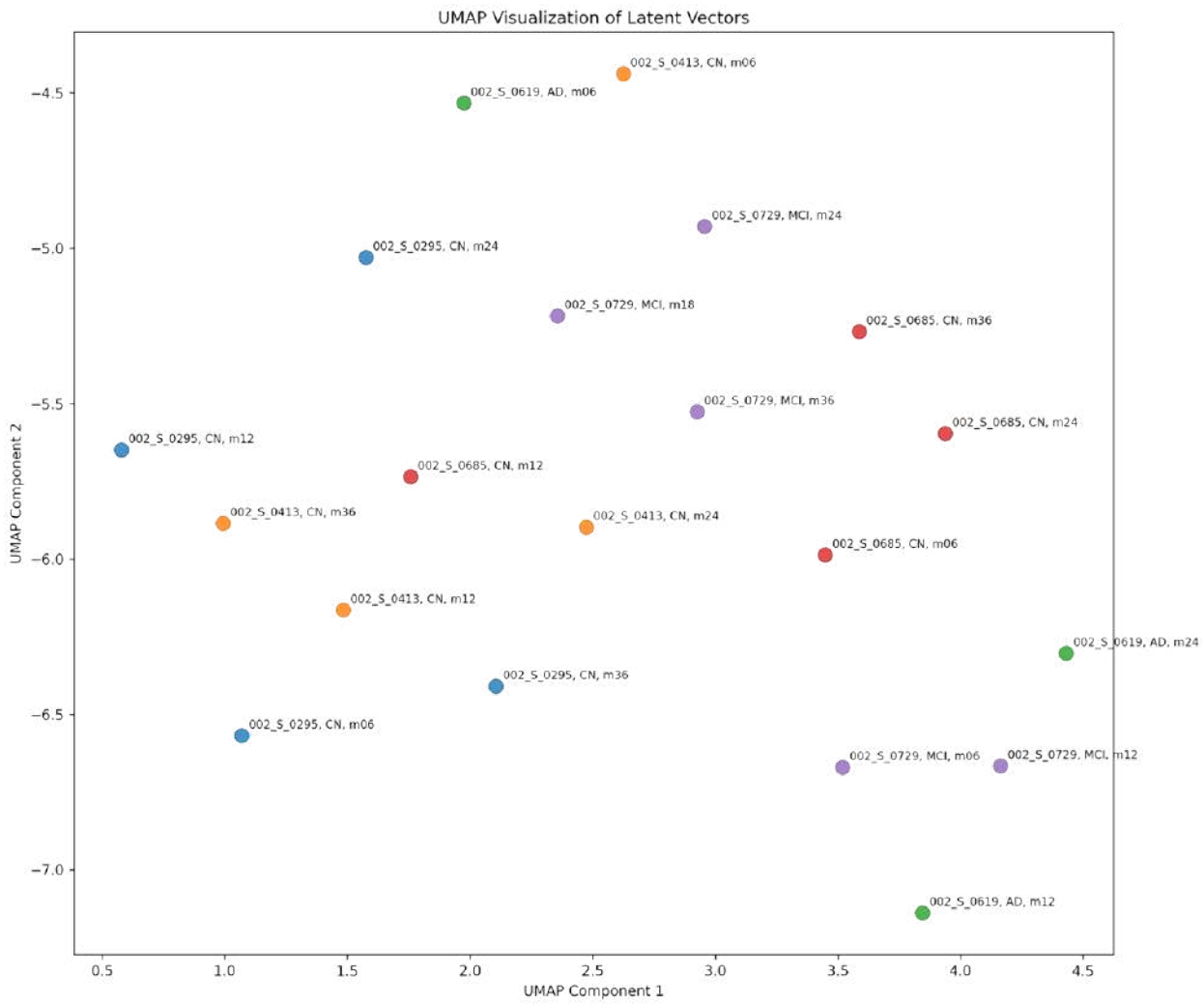


Figure 3.3: Visualization of latent space structure in Dense Augmented variant which is trained with a larger set of augmentations. This model. trained with more augmentations, shows slightly more localized clustering and improved intra-subject grouping.

conditioning during training, guiding the VAE to disentangle temporal progression from static anatomical variation.

3.4.1 Simple Time-Conditioned VAE

In this variant, we introduce explicit temporal supervision by conditioning the latent space on timepoint information. Specifically, we learn discrete timepoint embeddings that are fused with the latent representation using a cross-attention mechanism. The fused embeddings are then passed through a classification head to predict the associated timepoint. A cross-entropy loss is applied to these predicted time logits, encour-

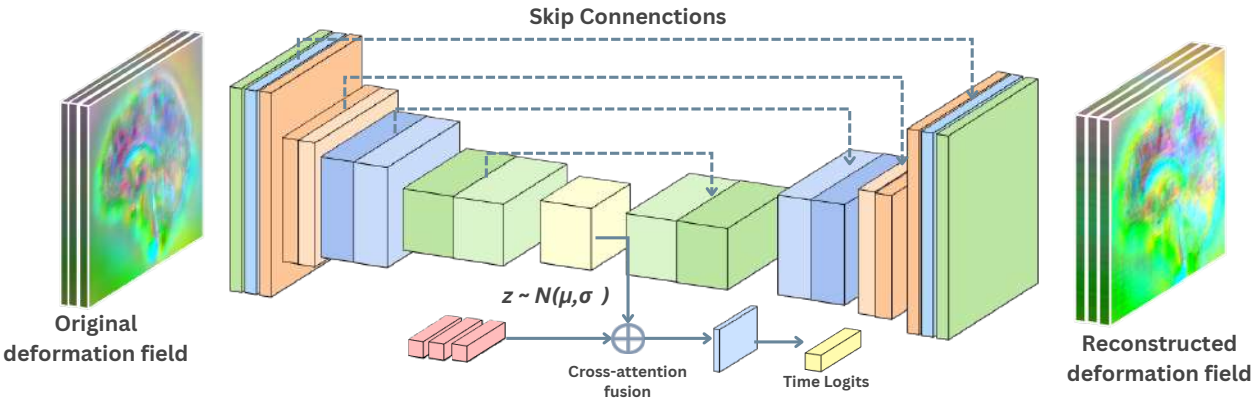


Figure 3.4: Architecture of the timepoint-aware transformer. A 3D stationary velocity field (SVF) is encoded into a latent representation. A learnable timepoint embedding is retrieved and fused with the latent representation via cross-attention. This fused latent is passed through a classification layer to obtain time logits. The fused representation is used to reconstruct the SVF while the time logits are used to classify the acquisition timepoint, encouraging temporally structured latent encoding. The trained model is then used in a downstream transformer to predict subject-specific latent trajectories across time.

aging the model to structure its latent space in accordance with temporal progression. The architecture is illustrated in Figure 3.4.

Despite this supervision, the learned latent space remains compressed and lacks meaningful temporal structure. As shown in Table 3.3, the model achieves only marginal improvements in classification and ordering performance. A qualitative visualization of the latent space for 30 subjects is presented in Figure 3.5, which highlights the limited separation across timepoints and the absence of smooth latent trajectories.

Metric	**Value**
Test Loss	0.4878
Average End-Point Error (EPE)	0.0668
Average Angular Error (degrees)	10.93
Time Prediction Accuracy	0.2552
Timepoint Pairwise Ordering Accuracy	0.4865
Timepoint Classification Accuracy	0.1587
Timepoint F1 Score (Macro)	0.1254
Timepoint F1 Score (Weighted)	0.1494
Kendall's τ	0.5585
Spearman's ρ	0.5896

Table 3.3: Evaluation results for the Simple Time-Conditioned VAE trained with 16k synthetic samples. Despite direct time supervision, the model exhibits lower classification and ordering accuracy than the Dense Augmented VAE.

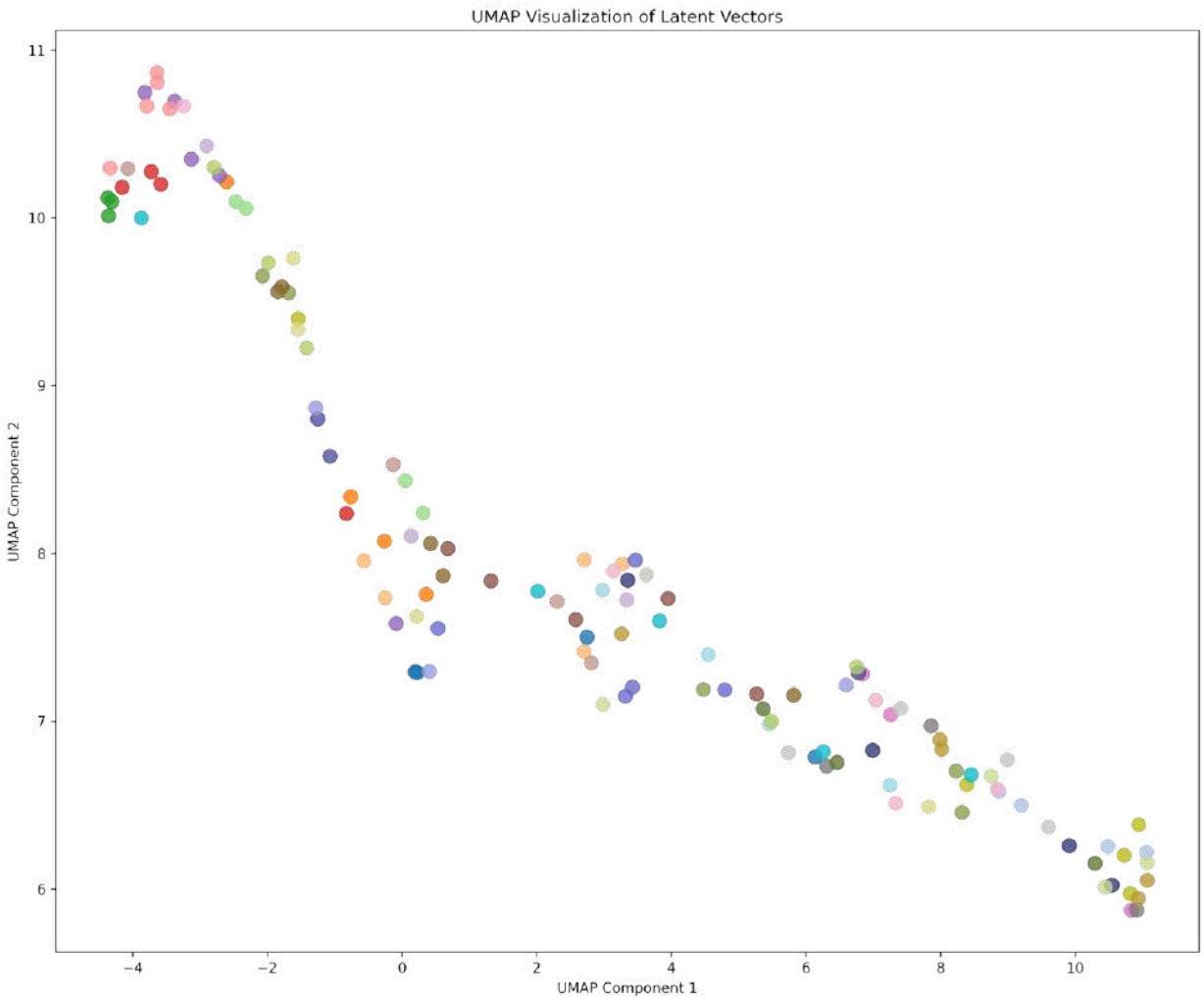


Figure 3.5: Latent space visualization for 30 subjects using a Simple Time-Conditioned VAE trained with 16k synthetic samples. Each color represents a unique subject. Subject-level clustering is more pronounced; however, the overall structure is compressed along a diagonal axis, limiting temporal resolution.

3.4.2 Factorized Time-Conditioned VAE

While classifier-based supervision provides a weak signal for temporal structure, it does not explicitly disentangle temporal dynamics from subject-specific anatomical variation. To address this, we introduce a factorized latent space architecture that separates each SVF's latent representation into two distinct components: a content embedding Z_c, which encodes static anatomical structure, and a time embedding Z_t, which captures temporal progression.

An overview of the architecture is provided in Figure 3.6. To enforce disentanglement between the two components, we introduce an independence loss that penalizes cosine similarity between Z_c and Z_t, encouraging them to encode orthogonal aspects of variation (illustrated in Figure 3.10(a)). This objective is

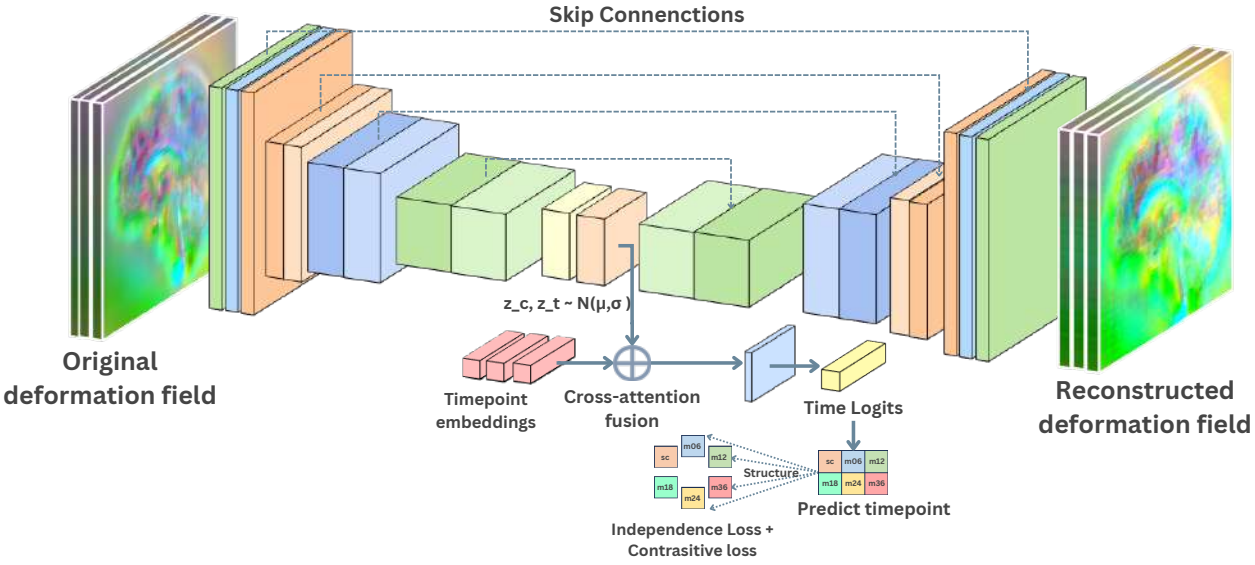


Figure 3.6: Overview of the architecture of the Factorized Timepoint-Aware Variational Autoencoder (FactorizedTimeVAE). A 3D stationary velocity field (SVF) is encoded into a latent representation, which is factorized into two distinct components: a content embedding Z_c that captures subject-specific anatomical structure, and a temporal embedding Z_t that models progression over time. An independence loss is applied to enforce disentanglement between Z_c and Z_t, ensuring that temporal and structural information are represented separately. A contrastive loss is applied to Z_t to encourage a temporally structured latent space aligned with chronological ordering.

mathematically defined as:

$$\mathcal{L}_{\text{indep}} = \left| \frac{1}{N} \sum_{i=1}^{N} \cos(\mathbf{z}_t^{(i)}, \mathbf{z}_c^{(i)}) \right|$$

This factorized model leads to improved temporal organization and latent interpretability, as evidenced by quantitative metrics (Tables 3.4 through 3.7). A qualitative comparison of the time and content components in latent space is shown in Figures 3.7 through 3.9, which highlights the enhanced subject clustering and partial separation of timepoints.

Metric	**Value**
Test Loss	0.3076
Average End-Point Error (EPE)	0.1015
Average Angular Error (degrees)	15.54
Time Prediction Accuracy	1.0000
Independence Loss (Cosine Similarity)	0.0478

Table 3.4: Test-time performance metrics for the FactorizedTimeVAE model. These values summarize reconstruction quality, directional consistency, and latent disentanglement.

Latent Component	Pairwise Accuracy	Classification Accuracy	F1 Macro / Weighted
Full Latent Space	0.5068	0.1619	0.1368 / 0.1568
Time Component	0.4842	0.1937	0.1716 / 0.1897
Content Component	0.5158	0.2444	0.2099 / 0.2306

Table 3.5: Comparison of classification and pairwise ordering performance across latent representations. The content component yields the highest classification accuracy and F1 scores, while the full latent space provides more balanced performance across metrics.

Metric	Full Latent Space	Time Component	Content Component
Timepoint Pairwise Ordering Accuracy	0.5068	0.4842	0.5158
Classification Accuracy	0.1619	0.1937	0.2444
F1 Score (Macro)	0.1368	0.1716	0.2099
F1 Score (Weighted)	0.1568	0.1897	0.2306

Table 3.6: Detailed breakdown of ordering and classification performance using different components of the FactorizedTimeVAE latent space. The content embedding achieves the best overall classification metrics, while pairwise ordering accuracy remains strongest for the content component.

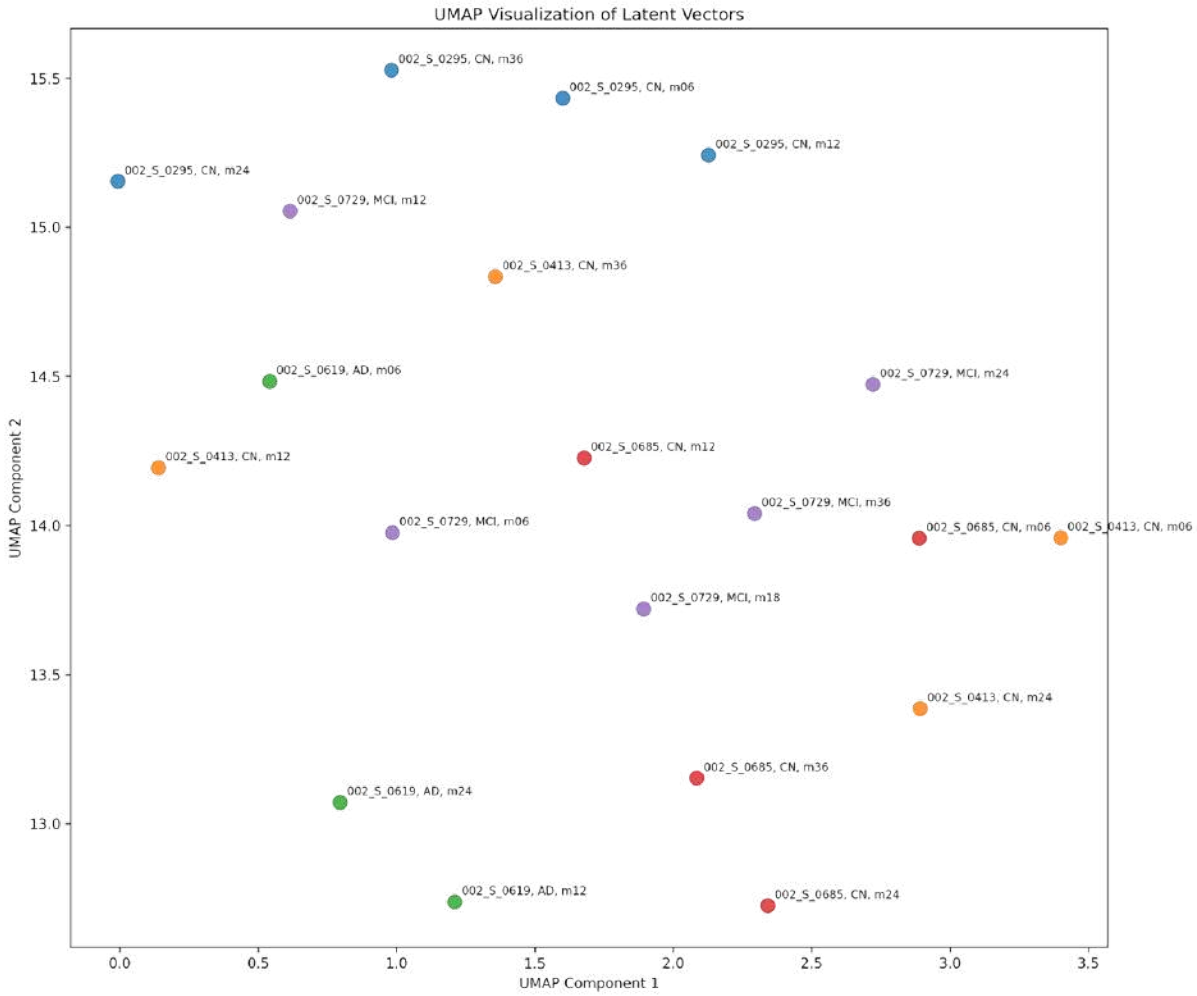
Latent Component	Kendall's τ	Spearman's ρ
Full Latent Space	0.5515	0.5714
Time Component	0.4647	0.5234
Content Component	0.4687	0.4948

Table 3.7: Rank correlation coefficients between predicted and true timepoints for different latent components. The full latent representation exhibits the strongest temporal alignment based on Kendall's τ and Spearman's ρ.

3.4.3 Contrastive Factorized Time-Conditioned VAE

While factorizing the latent space into anatomical content (Z_c) and temporal progression (Z_t) supports disentanglement, it does not by itself guarantee that Z_t reflects chronological order or forms a temporally coherent structure. To address this, we incorporate two key mechanisms: conditional priors for temporal structure, and a contrastive loss applied to the time-specific embedding.

Conditional Priors for Temporal Structure. Instead of assuming an isotropic Gaussian prior over Z_t, we define learnable priors for each discrete timepoint $t \in \{0, \ldots, T-1\}$. Each prior is parameterized by

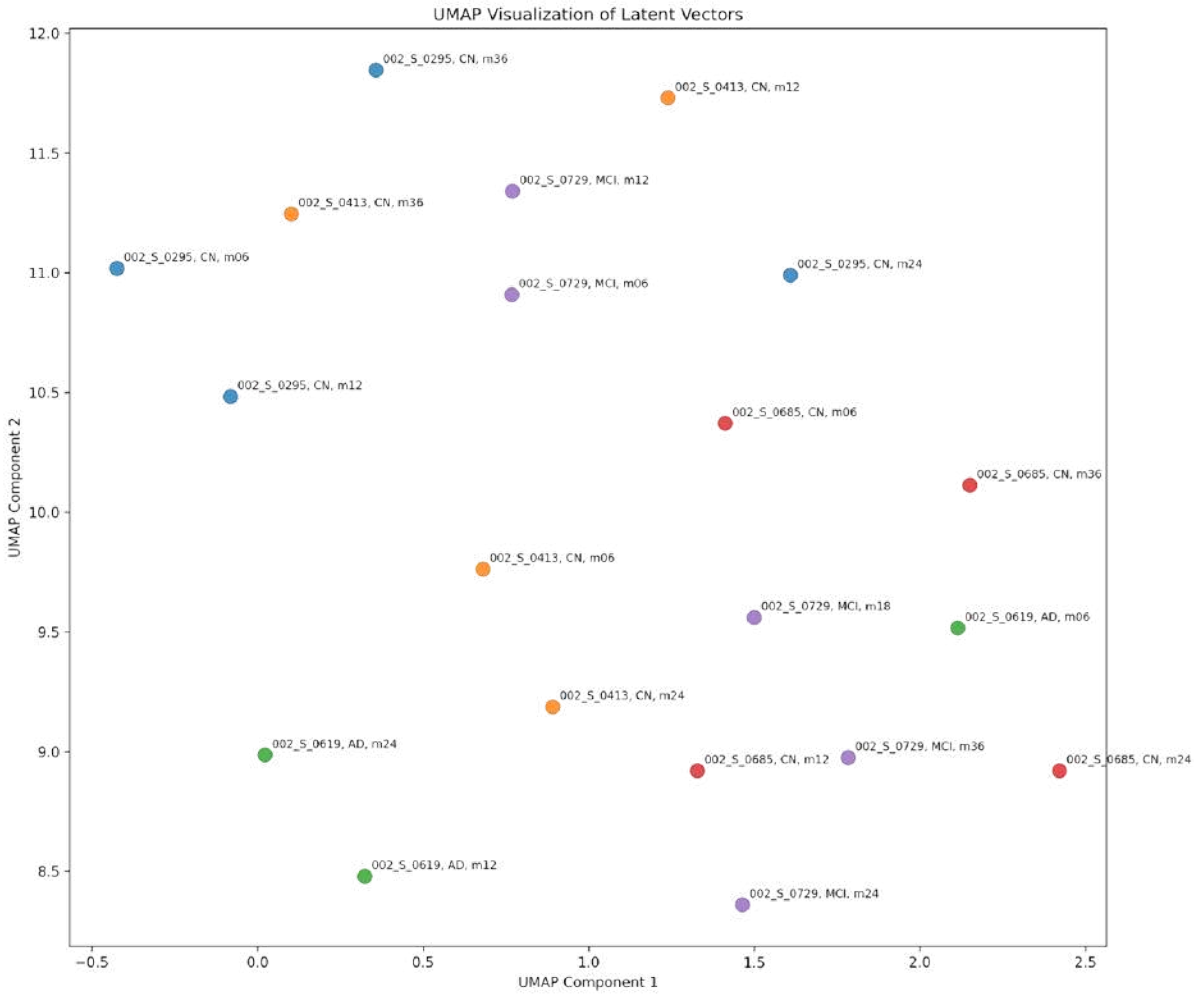


Full Component Latent Space

Figure 3.7: Latent space visualization of the full embedding in the *FactorizedTimeVAE* model trained without contrastive loss. This embedding combines both time and content components, but does not exhibit distinct clustering or structured temporal organization.

a mean $\mu_{\text{prior}}(t)$ and log-variance $\log \sigma^2_{\text{prior}}(t)$, implemented as embedding layers. During training, the KL divergence is computed between the approximate posterior $q(Z_t|x)$ and the corresponding time-conditioned prior $p(Z_t|t)$. This encourages embeddings to align with canonical representations of their respective timepoints, anchoring Z_t within a temporally meaningful coordinate system.

Contrastive Loss for Temporal Cohesion. To further enhance structure within the time-specific embedding, we apply a contrastive loss that encourages within-timepoint clustering while separating samples from



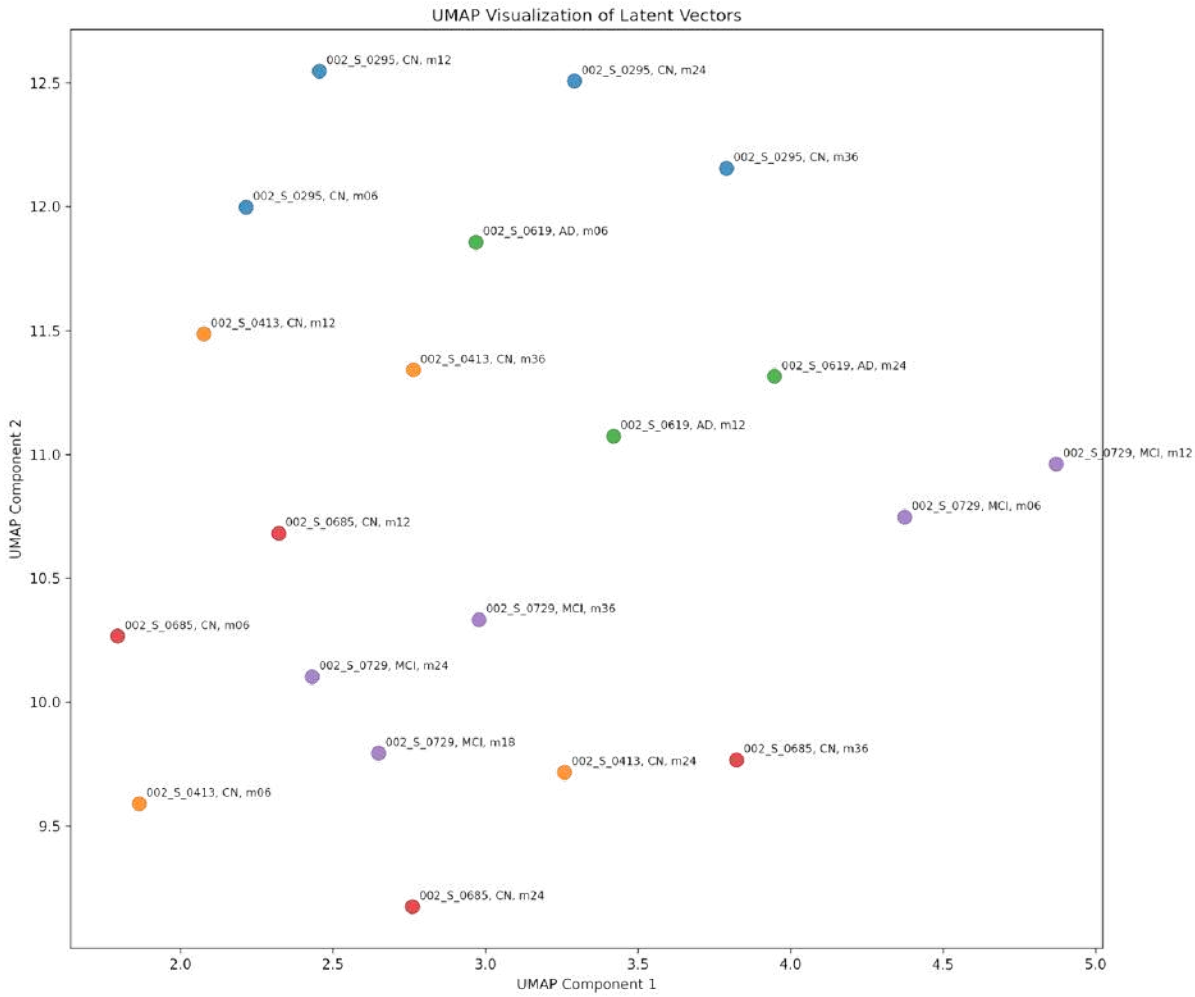
Time Component Latent Space

Figure 3.8: Latent space visualization of the time-specific component of the *FactorizedTimeVAE*. Compared to other embeddings, it exhibits clearer subject-level clustering, although a structured temporal arrangement is still lacking.

different timepoints. The objective is defined as:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{j\in\mathcal{P}(i)}\exp\left(\frac{\text{sim}(\mathbf{z}_t^{(i)},\mathbf{z}_t^{(j)})}{\tau}\right)}{\sum_{k\neq i}\exp\left(\frac{\text{sim}(\mathbf{z}_t^{(i)},\mathbf{z}_t^{(k)})}{\tau}\right)}$$

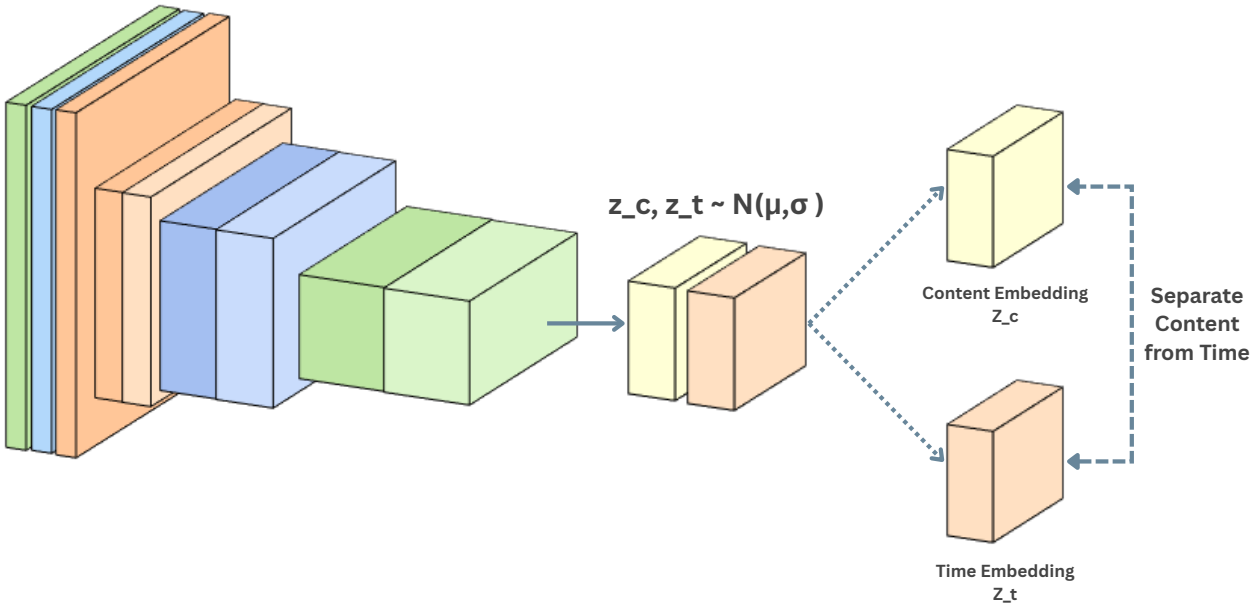
Here, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter, and $\mathcal{P}(i)$ indexes samples with the same timepoint label as i. This loss enforces local consistency and separability in the latent manifold, aligning anatomical variation with chronological progression. The contrastive loss formulation is illustrated in Figure 3.10(b).



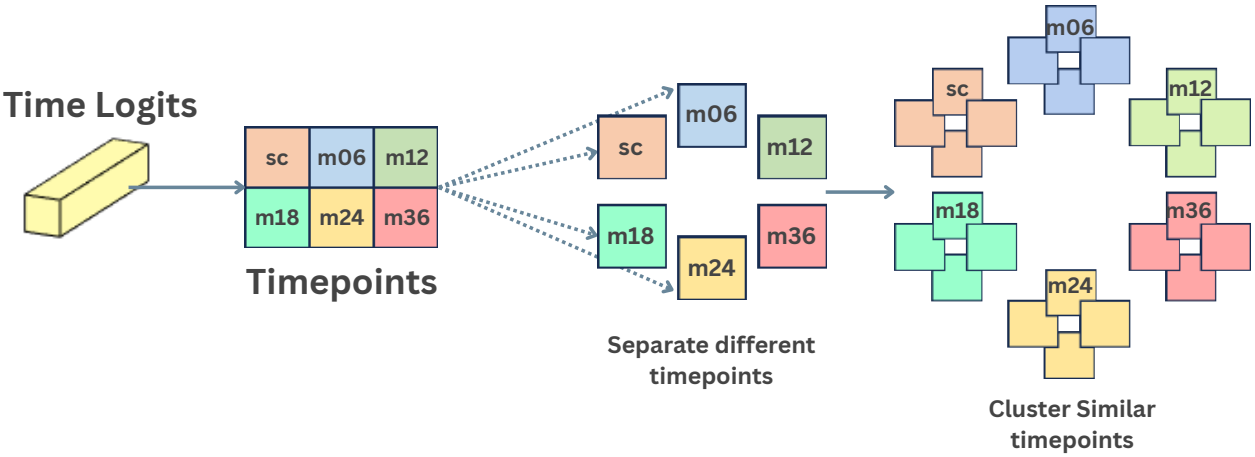
Content Component Latent Space

Figure 3.9: Latent space visualization of the content-specific component of the *FactorizedTimeVAE*. This space exhibits distinct clustering by subject but lacks any evident temporal structure.

Effect on Latent Space Organization. The combination of conditional priors and contrastive supervision yields a latent space that is both temporally disentangled and clinically interpretable. As shown in Figure 3.13, the representations for each subject form smooth, low-dimensional manifolds in the latent space, suggesting an implicit encoding of temporal progression. However, while intra-subject trajectories are coherent, the global ordering of timepoints is not strictly monotonic or chronological. This improvement in latent structure is achieved with minimal compromise to reconstruction fidelity or anatomical consistency. An important caveat of using UMAP for visualization is that, as a non-linear projection, it does not guarantee faithful preservation of temporal dynamics that may exist in the higher-dimensional latent space. A decomposition of the content and time-specific embeddings is provided in Figures 3.12 through 3.14, and the overall performance is corroborated by quantitative results in Tables 3.8–3.11. Figure 3.11 visualizes the latent space for 30 subjects where each color represents a unique subject.



((a)) Overview of the independence loss used in the FactorizedTimeVAE architecture. This loss penalizes cosine similarity between the time embedding $\mathbf{z}_t$ and the content embedding $\mathbf{z}_c$, encouraging them to encode statistically independent (orthogonal) information. As a result, the model disentangles temporal progression from subject-specific anatomical structure.



((b)) Overview of the contrastive loss applied to the time embedding $\mathbf{z}_t$ in the FactorizedTimeVAE model. "sc" represents the baseline/screening timepoint and "mX" represents the month X timepoint. This loss encourages samples from the same timepoint class to be closer in the latent space, while pushing apart dissimilar ones. This results in a temporally structured embedding space aligned with chronological relationships.

Figure 3.10: Illustration of the loss components used to structure the latent space in FactorizedTimeVAE. The independence loss promotes disentanglement between content and time embeddings, while the contrastive loss enforces temporal consistency in the time embedding space.

* Is Contrastive Better than Non-contrastive? A direct comparison of the time-specific latent embeddings between the contrastive and non-contrastive variants is shown in Figure 3.15. The contrastive model

Metric	Value
Test Loss	0.0019
Average End-Point Error (EPE)	0.0641
Average Angular Error (degrees)	11.36
Time Prediction Accuracy	1.0000
Independence Loss (Cosine Similarity)	0.0481

Table 3.8: Updated test-time performance metrics for the FactorizedTimeVAE model trained with contrastive loss. These results reflect improved reconstruction accuracy and latent disentanglement, as indicated by reduced EPE, angular error, and cosine similarity loss.

Latent Component	Pairwise Accuracy	Classification Accuracy	F1 Macro / Weighted
Full Latent Space	0.5090	0.1619	0.1261 / 0.1478
Time Component	0.5113	0.2222	0.1961 / 0.2192
Content Component	0.4932	0.1841	0.1665 / 0.1817

Table 3.9: Comparison of classification and pairwise ordering performance across latent representations for the factorized embeddings trained with contrastive loss. The time component shows the strongest performance, highlighting its relevance for modeling disease progression.

Metric	Full Latent Space	Time Component	Content Component
Timepoint Pairwise Ordering Accuracy	0.5090	0.5113	0.4932
Classification Accuracy	0.1619	0.2222	0.1841
F1 Score (Macro)	0.1261	0.1961	0.1665
F1 Score (Weighted)	0.1478	0.2192	0.1817

Table 3.10: Detailed breakdown of ordering and classification performance using different components of the FactorizedTimeVAE model trained with contrastive loss latent space. The time-specific latent variables contribute most effectively to temporal structure modeling.

demonstrates markedly better temporal continuity and subject-level coherence, supporting the quantitative improvements discussed earlier.

3.5 Results and Summary

We summarize the performance of all model variants across reconstruction accuracy, motion estimation, classification, and temporal consistency metrics. Each variant reflects a progression in architectural design and training strategy, culminating in the final Contrastive Factorized Time-Conditioned VAE.

Latent Component	Kendall's τ	Spearman's ρ
Full Latent Space	0.5900	0.6000
Time Component	0.4860	0.5442
Content Component	0.4886	0.5416

Table 3.11: Rank correlation coefficients for predicted versus true timepoints across different latent components trained with contrastive loss. Full latent representations exhibit the strongest alignment with ground-truth temporal orderings.

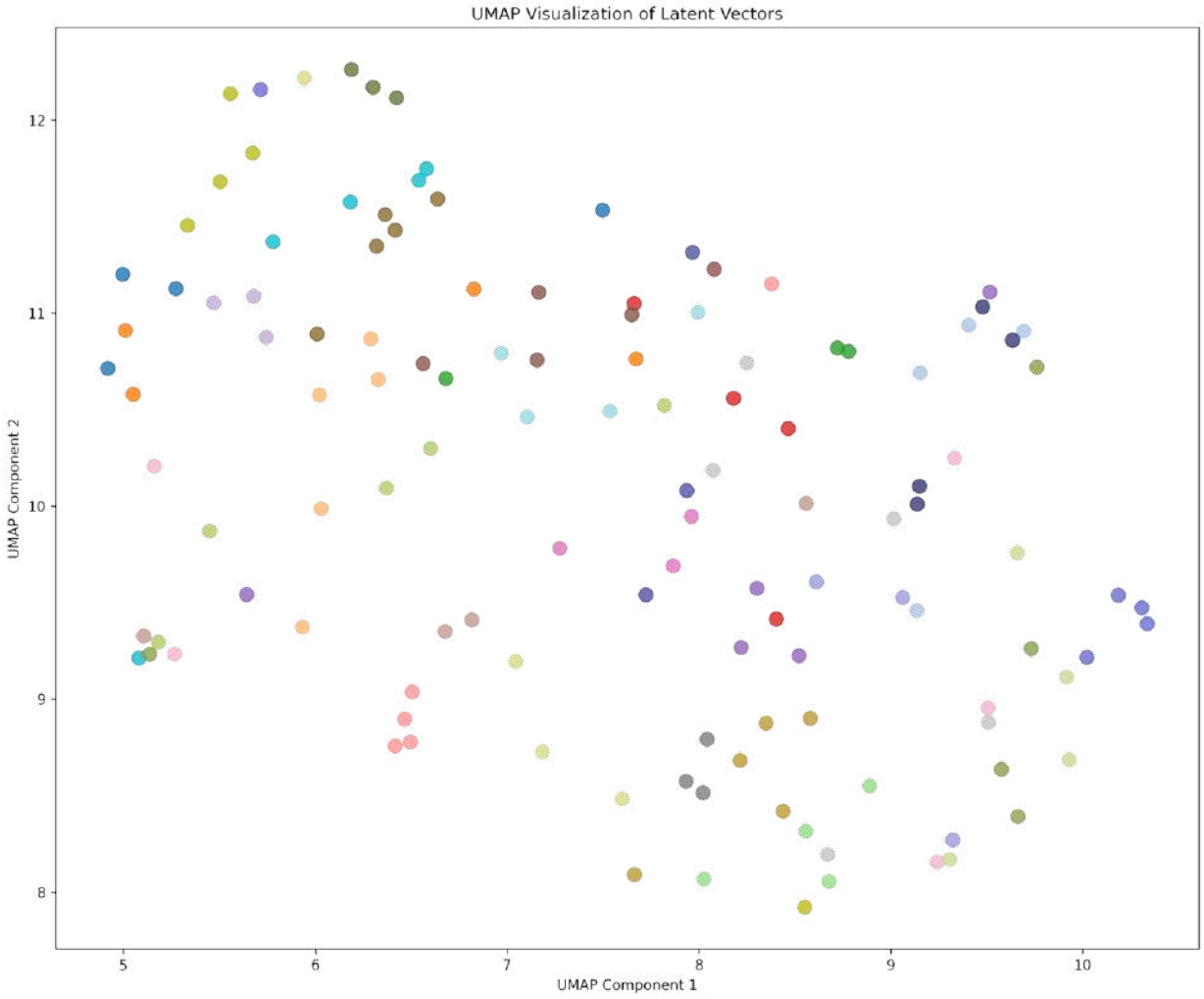
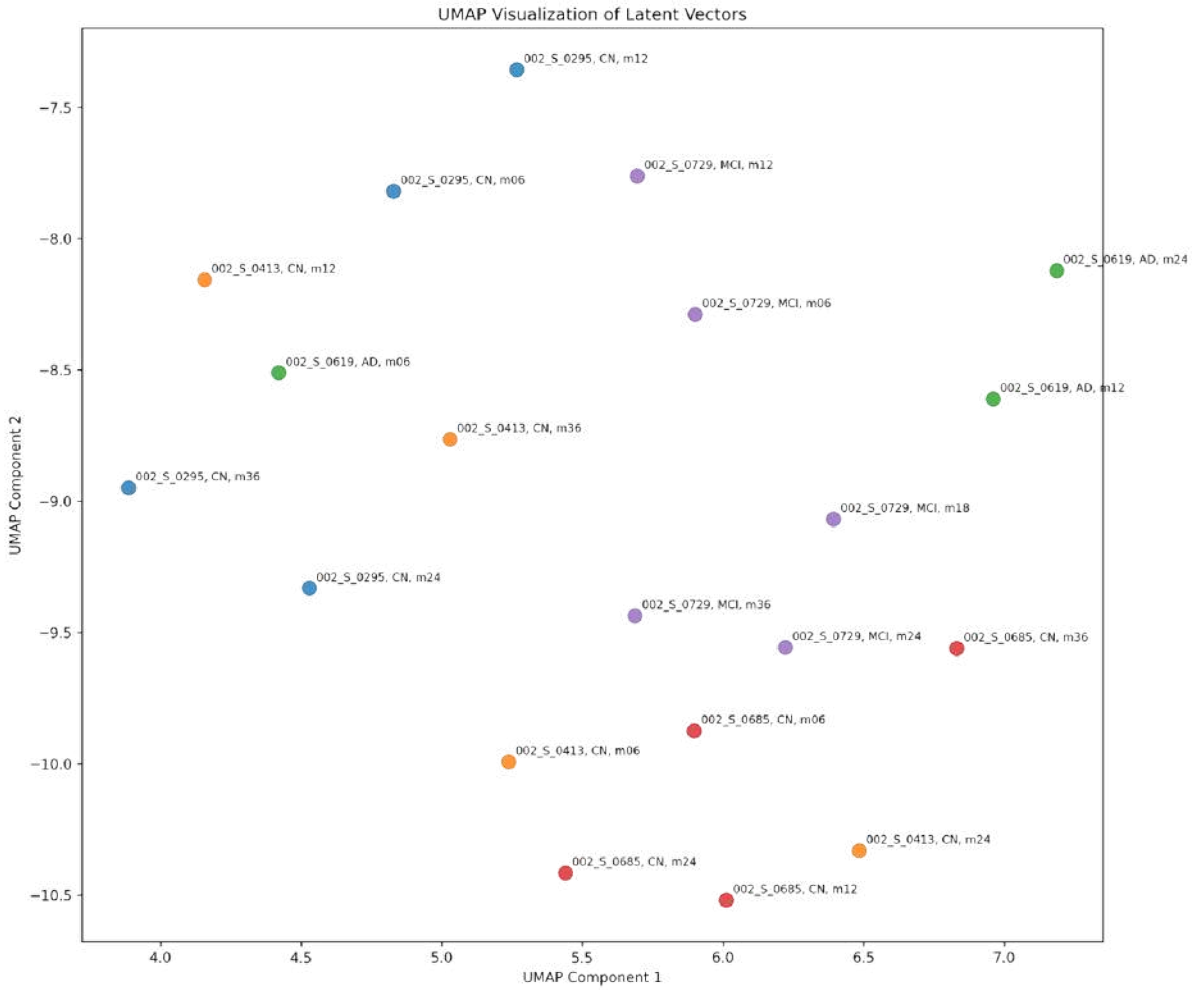


Figure 3.11: Latent space visualization for 30 subjects using a Factorized Time-Conditioned VAE trained with contrastive loss on 16k synthetic samples.

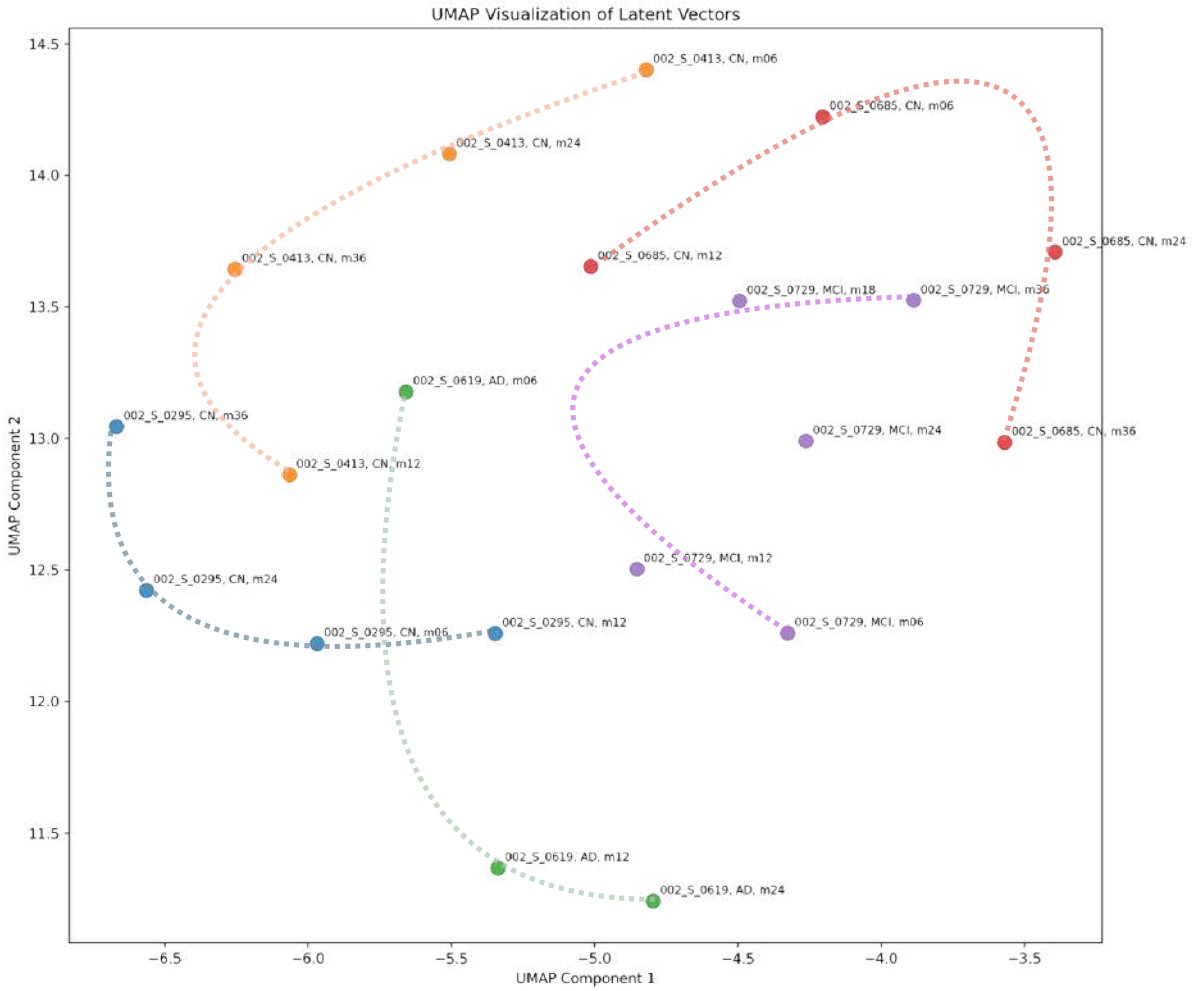


Full Component Latent Space

Figure 3.12: Latent space visualization of the full embedding in the *FactorizedTimeVAE* model trained with contrastive loss. This embedding integrates both time and content features and exhibits improved subject-wise separation and temporal continuity compared to the non-contrastive variant.

3.5.1 Quantitative Comparison

To assess the impact of augmentation and conditioning strategies across VAE variants, we evaluate each model on a suite of quantitative metrics that reflect both reconstruction accuracy and latent space structure. Table 3.12 reports the loss components used for all the variants and the corresponding test loss across all models. Interestingly, models conditioned on time or metadata do not necessarily achieve lower reconstruction loss. This suggests that increased supervision (e.g., classification or contrastive objectives) encourages better latent structure at the cost of slight reconstruction degradation.



Time Component Latent Space

Figure 3.13: Latent space visualization of the time-specific component of the *FactorizedTimeVAE* trained with contrastive loss. This representation shows the clearest clustering of subjects and the latent representations for each subject lie along smooth, continuous contours, highlighted by dotted lines.

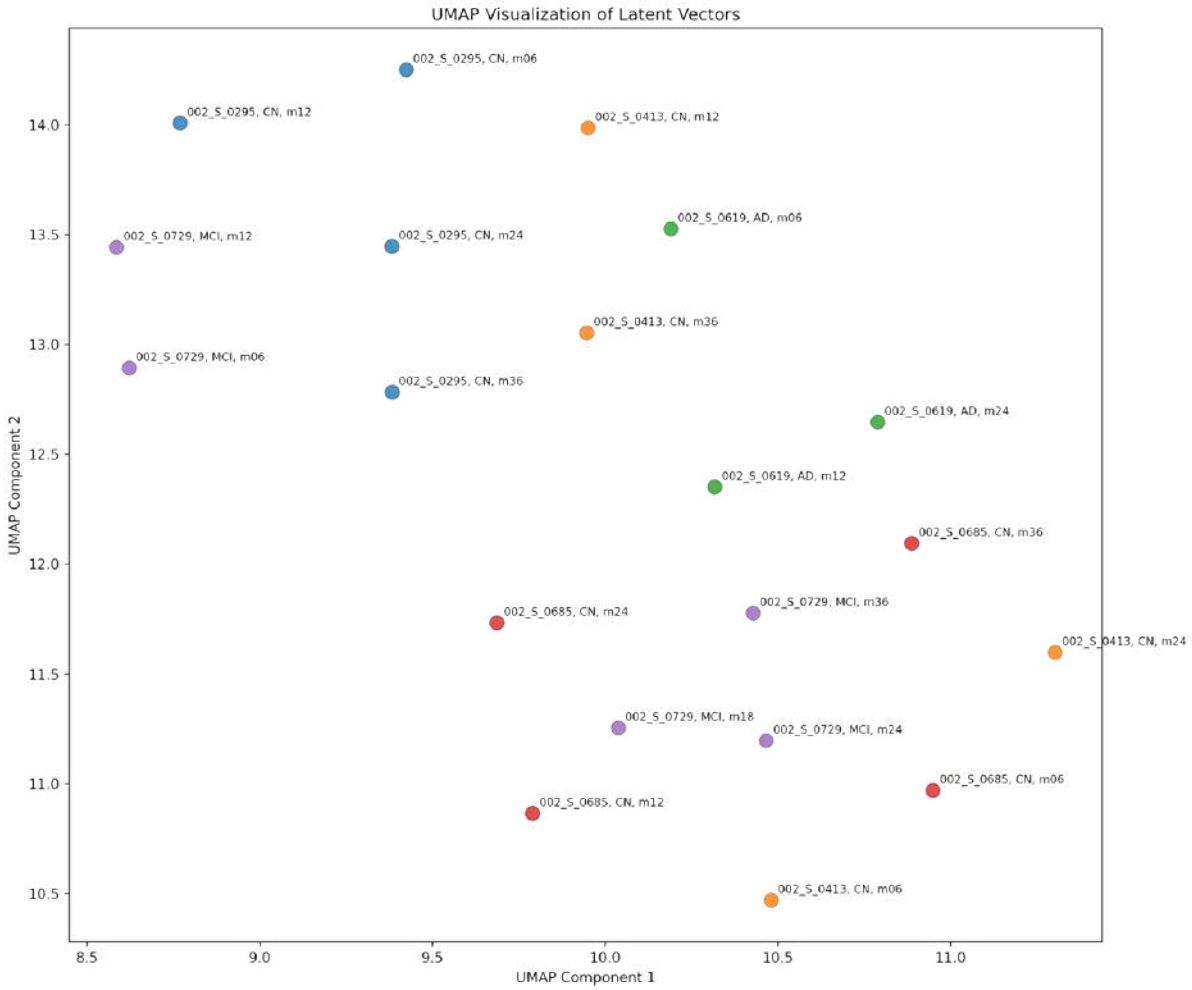
3.5.1.1 SVF Reconstruction Accuracy

To evaluate the fidelity of reconstructed stationary velocity fields (SVFs), we report two commonly used deformation metrics: End-Point Error (EPE) and Average Angular Error (AAE), shown in Table 3.13.

End-Point Error (EPE) measures the Euclidean distance between predicted and ground-truth vector fields at each voxel:

$$\text{EPE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{v}_i^{\text{pred}} - \mathbf{v}_i^{\text{gt}}\|_2$$

where $\mathbf{v}_i^{\text{pred}}$ and $\mathbf{v}_i^{\text{gt}}$ are the predicted and ground-truth SVF vectors at voxel i, and N is the total number of voxels.



Content Component Latent Space

Figure 3.14: Latent space visualization of the content-specific component of the *FactorizedTimeVAE* with contrastive loss. While subject-wise clustering is improved, the temporal structure remains less pronounced compared to the time-specific space.

Average Angular Error (AAE) evaluates the mean angular deviation between predicted and ground-truth deformation directions:

$$\text{AAE} = \frac{1}{N} \sum_{i=1}^{N} \arccos \left(\frac{\langle \mathbf{v}_i^{\text{pred}}, \mathbf{v}_i^{\text{gt}} \rangle}{\|\mathbf{v}_i^{\text{pred}}\|_2 \cdot \|\mathbf{v}_i^{\text{gt}}\|_2} \right)$$

Results show that the Vanilla VAE achieves the lowest EPE, indicating strong voxel-level accuracy. However, the Dense Augmented VAE achieves the best angular alignment, suggesting that structured augmentation improves directional coherence of deformations. In contrast, both the Simple Time-Conditioned and Factorized models exhibit degraded angular performance. This may reflect trade-offs introduced by addi-

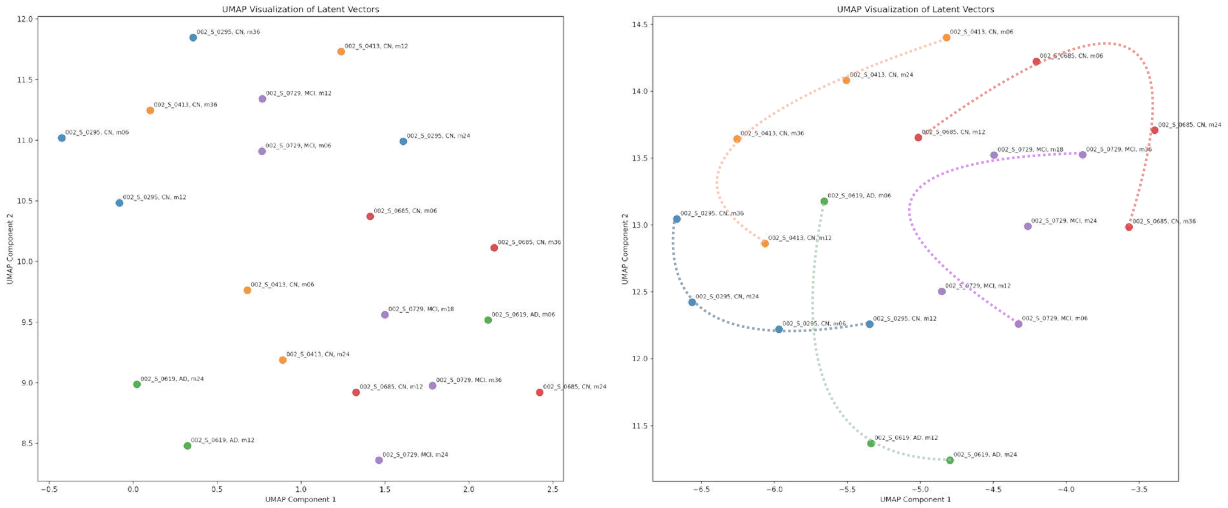


Figure 3.15: Comparison of the latent space structure of the time component in Factorized VAE models trained with and without contrastive loss. Both models use the time-specific embedding, but the variant trained with contrastive loss exhibits improved subject-level clustering and a more continuous temporal arrangement of timepoints along a low-dimensional manifold.

Model	**No. Training Samples**	**Age Cond.**	**Disease Cond.**	**Time Cond.**	**Loss Components**	**Best Validation Loss ↓**	**Test Loss ↓**
Vanilla VAE	2140 (1140 real, 1000 synthetic)	✗	✗	✗	MSE + KL Divergence	0.12229	0.2882
Dense Augmented VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✗	MSE + KL Divergence	0.12279	0.2891
Simple Time-conditioned VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	MSE + KL + CE(Time)	0.31157	0.4832
Non-contrastive Factorized VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	MSE + KL + CE(Time) + Independence	0.32214	0.3076
Contrastive Factorized VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	MSE + KL + CE(Time) + Independence + Contrastive	0.13333	0.0019

Table 3.12: Comparison of VAE models in terms of training data, conditioning strategies, loss functions, and performance. All models are trained to reconstruct SVFs, with additional loss components used in time-aware variants.

tional supervision—such as classification and disentanglement losses—which bias the model away from purely spatial accuracy in favor of more semantically structured latent representations.

3.5.1.2 Temporal Ordering and Classification Performance

To assess how well each model captures temporal progression in the latent space, we report timepoint classification accuracy, pairwise ordering accuracy, and two rank correlation metrics: Kendall's τ and Spearman's ρ (Table 3.14). Both τ and ρ measure the consistency of predicted timepoint orderings relative to the ground truth. While Spearman's ρ assesses monotonic relationships based on rank differences, Kendall's τ evaluates concordance by counting pairwise rank agreements, making it more robust to small rank shifts.

Model	No. Training Samples	Age Cond.	Disease Cond.	Time Cond.	Average EPE ↓	Average Angular Error (deg) ↓
Vanilla VAE	2140 (1140 real, 1000 synthetic)	✗	✗	✗	**0.0584**	8.0740
Dense Augmented VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✗	0.0639	**7.2464**
Simple Time-conditioned VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	0.0546	8.6180
Non-contrastive Factorized VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	0.1015	15.54
Contrastive Factorized VAE	16900 (1140 real, 15760 synthetic)	✗	✗	✓	0.0641	11.36

Table 3.13: Comparison of SVF reconstruction performance across VAE models using Average End-Point Error (EPE) and Average Angular Error. All values are computed on the predicted SVFs.

The results show that explicit temporal conditioning significantly improves downstream performance. The Contrastive Factorized VAE achieves perfect timepoint classification accuracy and the highest F1 scores, indicating that contrastive loss provides a strong supervisory signal for structuring the time-specific embedding. Although the Vanilla and Dense Augmented models perform surprisingly well in Kendall's τ and Spearman's ρ, these gains are not reflected in pairwise or classification accuracy, suggesting that their ordering is partially preserved but lacks semantic grounding.

Model	Time Accuracy	Pairwise Acc.	Cls Acc.	F1 Macro	F1 Weighted	Kendall Tau	Spearman Rho
Vanilla VAE	N/A	0.5113	0.1619	0.1396	0.1577	**0.5124**	**0.5675**
Dense Augmented VAE	N/A	**0.5315**	0.1683	0.1511	0.1700	0.5942	0.6182
Simple Time-conditioned VAE	0.2677	0.4392	0.1429	0.1220	0.1435	0.5980	0.6013
Non-contrastive Factorized VAE	**1.0000**	0.4842	0.1937	0.1716	0.1897	0.4647	0.5234
Contrastive Factorized VAE	**1.0000**	0.5113	**0.2222**	**0.1961**	**0.2192**	0.4860	0.5442

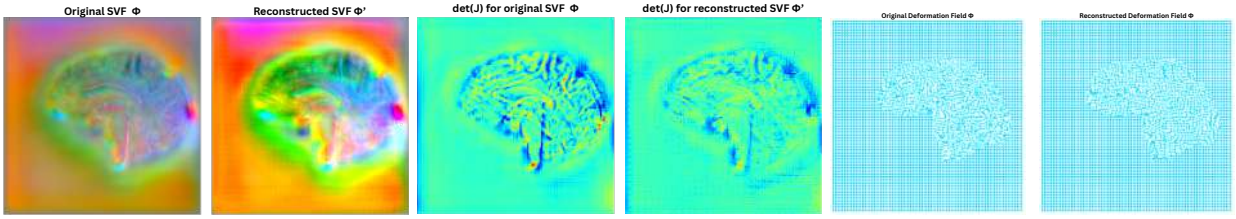
Table 3.14: Evaluation of VAE models on timepoint ordering and classification performance. Time prediction accuracy is only reported for time-conditioned models. For the factorized models, metrics are based on the time-specific embedding.

3.5.2 Qualitative Evaluation

Figures 3.16 through 3.20 visualize reconstructed SVFs, Jacobian determinants, and deformation grids for Subject 137 across all model variants.

- **Vanilla VAE** (Fig. 3.16): Reconstructs coarse deformation but fails to capture fine details, especially in grid warping.

- **Dense Augmented VAE** (Fig. 3.17): Shows improved alignment and smoother Jacobians due to synthetic augmentation.
- **Simple Time-Conditioned VAE** (Fig. 3.18): Latent guidance does not yield visible gains in deformation fidelity, indicating weak conditioning.
- **Factorized VAE** (Fig. 3.19): Deformations appear sharper, but latent disentanglement introduces some loss in reconstruction accuracy.
- **Contrastive Factorized VAE** (Fig. 3.20): Best balances spatial smoothness and deformation plausibility while ensuring well-separated latent components.

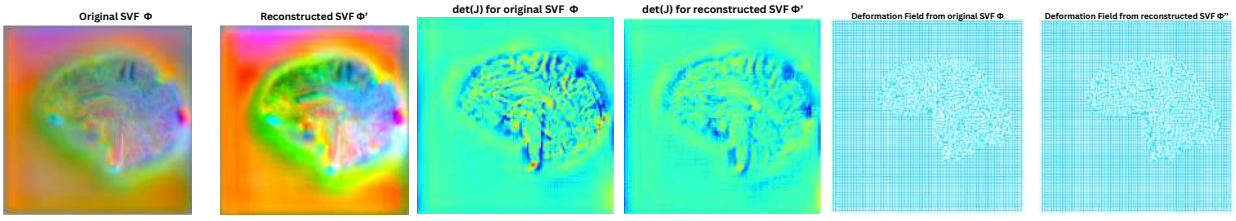


((a)) A comparision of the reconstruction quality of the SVF. Here, the SVF is RGB colored using each channel (dx, dy, dz)

((b)) Determinant of Jacobian visualization.

((c)) A visualization of how the deformation field warps a grid. The deformation is restricted to the brain region via the subject's brain mask.

Figure 3.16: A side by side comparision of original and reconstructed Stationary velocity fields for the Vanilla VAE Model with only 1k synthetic augmentations and no conditioning. All visualizations are shown for a middle sagittal slice for Subject 137.



((a)) A comparision of the reconstruction quality of the SVF. Here, the SVF is RGB colored using each channel (dx, dy, dz)

((b)) Determinant of Jacobian visualization.

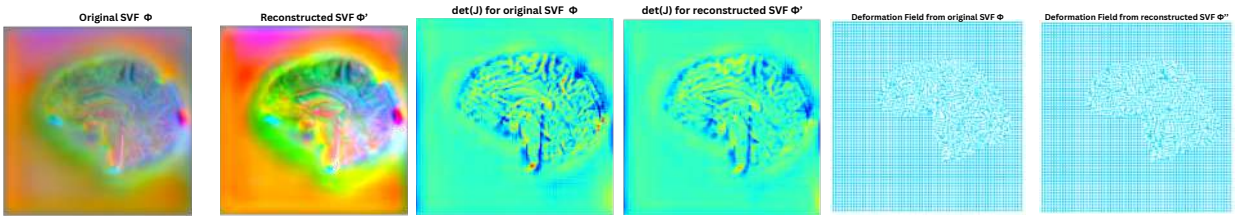
((c)) A visualization of how the deformation field warps a grid. The deformation is restricted to the brain region via the subject's brain mask.

Figure 3.17: A side by side comparision of original and reconstructed Stationary velocity fields for the Dense Augemnted VAE Model with 16k synthetic augmentations. All visualizations are shown for a middle sagittal slice for Subject 137.

3.5.3 Summary

In summary, the following trends emerge:

- Dense augmentation improves spatial alignment but does not meaningfully organize latent space.
- Simple time-conditioning via classifier loss introduces weak temporal structure.

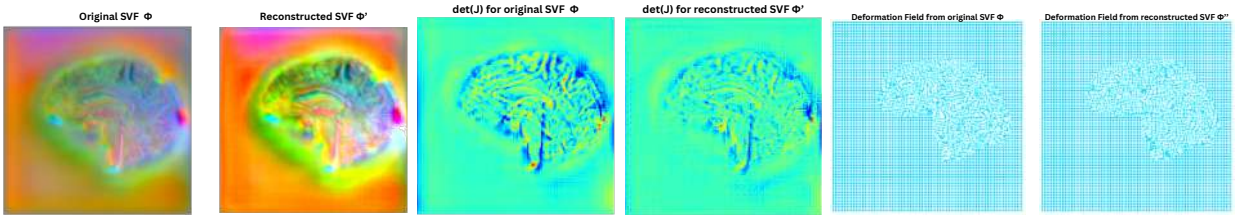


((a)) A comparision of the reconstruction quality of the SVF. Here, the SVF is RGB colored using each channel (dx, dy, dz)

((b)) Determinant of Jacobian visualization.

((c)) A visualization of how the deformation field warps a grid. The deformation is restricted to the brain region via the subject's brain mask.

Figure 3.18: A side by side comparision of original and reconstructed Stationary velocity fields for the Simple Time Aware VAE Model with 16k synthetic augmentations. All visualizations are shown for a middle sagittal slice for Subject 137.

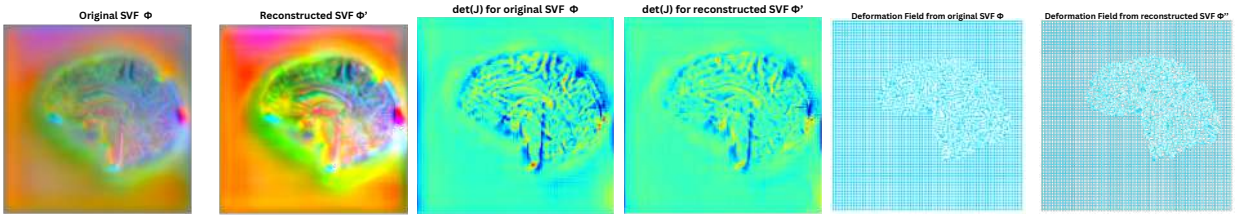


((a)) A comparision of the reconstruction quality of the SVF. Here, the SVF is RGB colored using each channel (dx, dy, dz)

((b)) Determinant of Jacobian visualization.

((c)) A visualization of how the deformation field warps a grid. The deformation is restricted to the brain region via the subject's brain mask.

Figure 3.19: A side by side comparision of original and reconstructed Stationary velocity fields for the Factorized Time Aware VAE Model with 16k synthetic augmentations. All visualizations are shown for a middle sagittal slice for Subject 137.



((a)) A comparision of the reconstruction quality of the SVF. Here, the SVF is RGB colored using each channel (dx, dy, dz)

((b)) Determinant of Jacobian visualization.

((c)) A visualization of how the deformation field warps a grid. The deformation is restricted to the brain region via the subject's brain mask.

Figure 3.20: A side by side comparision of original and reconstructed Stationary velocity fields for the contrastive Factorized Time Aware VAE Model with 16k synthetic augmentations. All visualizations are shown for a middle sagittal slice for Subject 137.

- Factorized embeddings enable better disentanglement but require explicit losses for structure.
- Contrastive supervision on the temporal component leads to the most coherent latent trajectories and best downstream performance.

These results validate the design principles of DIVA, demonstrating that combining metadata guidance with architectural disentanglement and contrastive supervision yields clinically meaningful, temporally consistent latent representations.

3.5.4 Metadata Conditioning for Latent Structure

In neuroimaging datasets, *metadata* refers to auxiliary information about each scan that is not part of the raw imaging data itself. Common examples include patient demographics (e.g., age, sex), diagnostic labels (e.g., cognitively normal (CN), mild cognitive impairment (MCI), Alzheimer's disease (AD)), and acquisition-specific variables such as timepoint or scanner type. While metadata is often discarded during representation learning, it encodes crucial context about the subject's clinical and temporal state.

In the case of progressive neurodegenerative diseases such as Alzheimer's, the same anatomical structure can differ significantly in appearance depending on both the disease stage and the subject's age. Conditioning latent representations on metadata provides a way to inject domain knowledge into the model and encourage disentangled representations that reflect clinically meaningful axes of variation [68, 69].

Without such conditioning, VAEs often entangle anatomical and temporal factors into a single latent vector, resulting in representations that are difficult to interpret and poorly aligned with clinical trajectories [61]. By contrast, metadata-informed VAEs can produce structured latent spaces that align with known disease progression patterns, enabling improved generalization and interpretability [70, 71]. In this work, we condition the VAE on two key metadata attributes: age (a continuous variable) and diagnosis (a categorical variable), encouraging the latent space to reflect both gradual anatomical aging and discrete diagnostic shifts.

3.5.5 Factorized Metadata-Conditioned VAE

While direct conditioning on metadata improves latent alignment with clinically relevant variables, it does not guarantee disentanglement of different sources of variation. In particular, age and diagnosis affect spatial deformations in distinct ways: age introduces smooth anatomical changes due to normal aging, while disease labels represent discrete pathology stages. To better capture these factors, we adopt a *factorized* latent architecture that separates metadata-driven information from subject-specific anatomical content.

In this variant, the VAE's latent space is split into two components: a **metadata embedding** Z_m conditioned on age and diagnosis, and a **content embedding** Z_c that captures residual anatomical variability independent of metadata. These two latent vectors are trained jointly, with the encoder predicting both branches. To promote disentanglement, an **independence loss** minimizes the cosine similarity between Z_m and Z_c, encouraging orthogonality and information separation.

This model is trained using a composite loss that includes: (1) standard VAE reconstruction and KL terms for both latent branches, (2) classification and regression losses for diagnosis and age prediction from Z_m, and (3) the independence loss. Table 3.15 summarizes these loss components for test and validation splits.

Split	Total Loss	Recon Loss	KLD Meta	KLD Content	Disease Loss	Age Loss	Independence Loss
Test	53.6237	0.2967	6.0840	0.0017	1.0557	288.8236	0.0145
Validation	55.9892	0.1266	5.9238	0.0068	1.1441	320.4394	0.0232

Table 3.15: Loss components of the metadata-conditioned VAE on test and validation sets.

Despite this structured design, the model does not achieve strong clustering by disease in the latent space. Figure 3.21 shows that embeddings from different diagnostic groups largely overlap. This is confirmed by the low silhouette score in Table 3.17(a). The overall disease classification accuracy remains moderate (Table 3.16), suggesting that the disentanglement helps isolate some metadata-driven variation, but does not yield fully separable class representations.

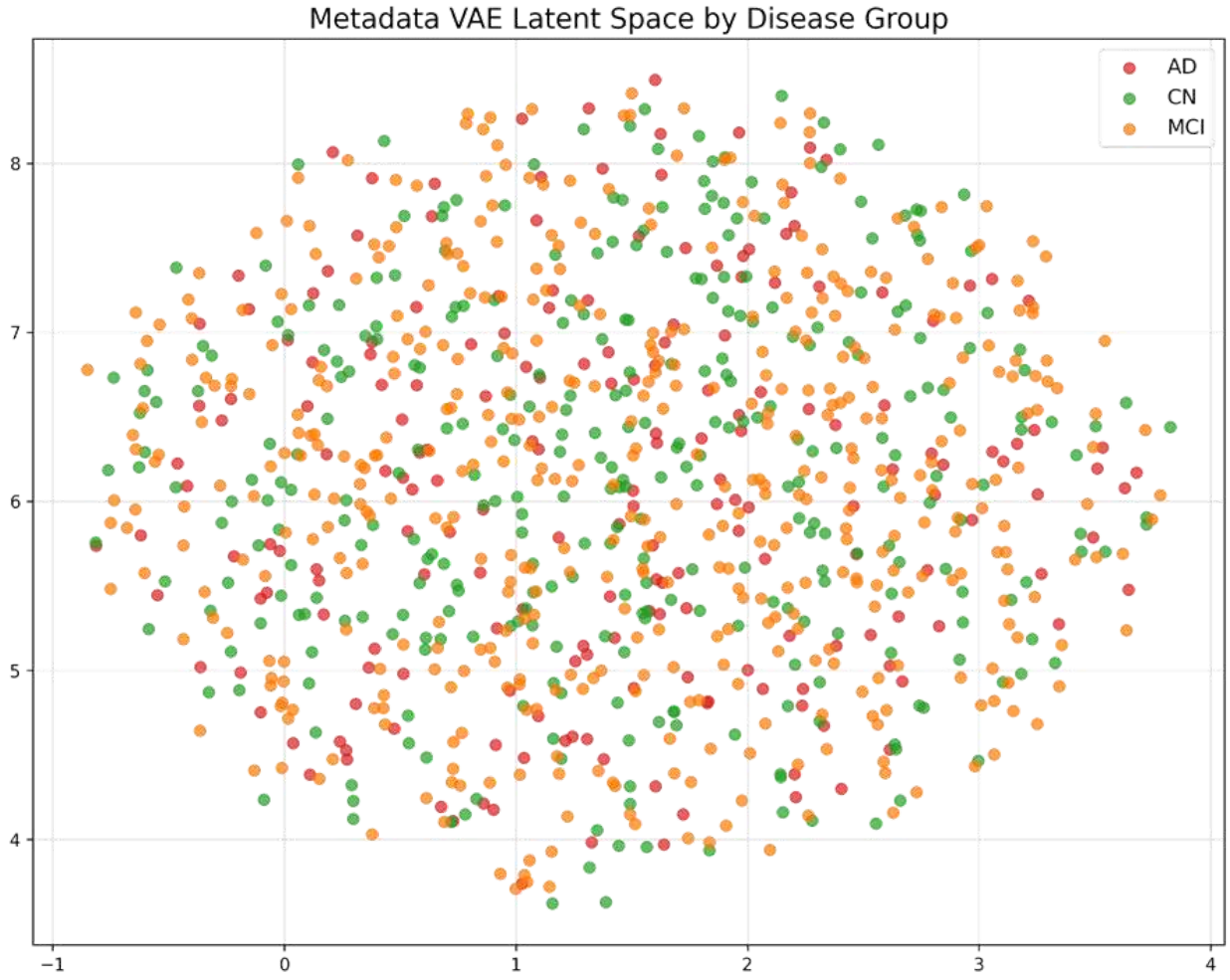


Figure 3.21: Latent space visualization for 100 samples using a Factorized Metadata-Conditioned VAE trained without contrastive loss on 16k synthetic samples. Despite explicit conditioning on metadata, the latent space is not well-separated as per disease label.

Split	Disease Prediction Accuracy
Test	0.4464
Validation	0.4097

Table 3.16: Evaluation metrics for the metadata-conditioned VAE model on test and validation sets.

Metric	Value
Silhouette Score	-0.0167
Number of Samples	1000

((a)) Clustering metrics for latent space of metadata-conditioned VAE.

Group	Count (%)
AD	198 (19.8%)
CN	330 (33.0%)
MCI	472 (47.2%)

((b)) Group distribution in clustered latent space.

Table 3.17: Latent space clustering evaluation using silhouette score and group distributions.

However, the model does demonstrate plausible anatomy-aware reconstructions and supports counterfactual reasoning. Figure 3.22 compares reconstructed SVFs across CN and MCI subjects with similar ages, showing that the model captures disease-driven deformation patterns. Figure 3.23 illustrates a counterfactual query where a CN subject is reconstructed as if they had AD, holding all other metadata constant.

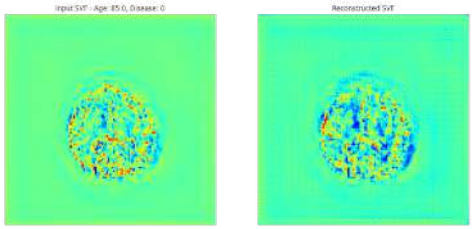
Overall, the factorized metadata-conditioned VAE introduces a structured latent design that promotes interpretability and enables clinically meaningful manipulation. While the DIVA architecture supports counterfactual querying, the reliability of such queries is limited by the latent space's expressiveness, which remains suboptimal. To address this, the next section introduces a contrastive training variant of the factorized metadata-conditioned VAE. This extension aims to enhance latent space organization and improve the semantic alignment of representations with metadata such as age and diagnosis.

3.5.6 Contrastive Factorized Metadata-Conditioned VAE

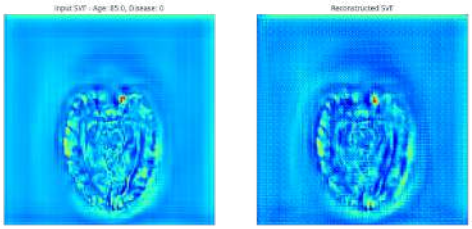
While the factorized metadata-conditioned VAE improves structural modularity by separating anatomical content from metadata embeddings, it does not consistently yield separable or interpretable latent clusters—particularly along the disease axis. To enhance structure and semantic alignment, we introduce a contrastive loss over the metadata embedding space. This is inspired by our earlier success with contrastive learning in temporal disentanglement (see Section 3.4.3).

Contrastive Loss for Metadata Alignment. To improve latent space separability with respect to clinical diagnosis, we define a contrastive loss over the metadata-specific embedding. Analogous to its temporal counterpart, this loss encourages embeddings from the same diagnosis group (e.g., AD, MCI, CN) to cluster together while pushing apart samples from different groups. The contrastive loss objective is given by:

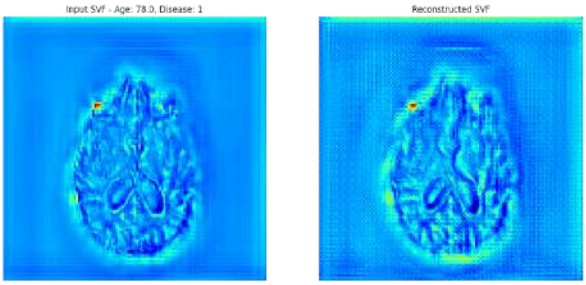
$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum\limits_{j\in\mathcal{P}(i)}\exp\left(\frac{\text{sim}(\mathbf{z}_m^{(i)},\mathbf{z}_m^{(j)})}{\tau}\right)}{\sum\limits_{k\neq i}\exp\left(\frac{\text{sim}(\mathbf{z}_m^{(i)},\mathbf{z}_m^{(k)})}{\tau}\right)}$$



((a)) Subject age: 85 years, diagnosis: cognitively normal (sample 1).



((b)) Subject age: 85 years, diagnosis: cognitively normal (sample 2).



((c)) Subject age: 78 years, diagnosis: mild cognitive impairment (MCI).

Figure 3.22: Comparison of original and reconstructed spatial vector fields (SVF) by the factorized metadata-conditioned VAE. The determinant of the Jacobian is visualized for each subject.

where $\mathbf{z}_m$ is the metadata embedding, $\mathcal{P}(i)$ denotes the set of samples sharing the same diagnosis as i, and τ is a temperature hyperparameter. This formulation encourages disease-wise alignment and enhances counterfactual consistency in the latent space.

Results and Limitations. Despite introducing the contrastive loss and conditional priors for disease and age metadata, latent space separability remains limited. As shown in Figure 3.24, neither the metadata nor content components exhibit strong clustering by diagnosis. This is further corroborated by clustering metrics (Table 3.18), where silhouette scores remain negative, indicating weak class separation.

Moreover, disease classification accuracy does not improve over the non-contrastive variant (Table 3.20), and in some cases, deteriorates on the validation set. While the total loss is reduced (Table 3.19), this does not translate into more discriminative embeddings.

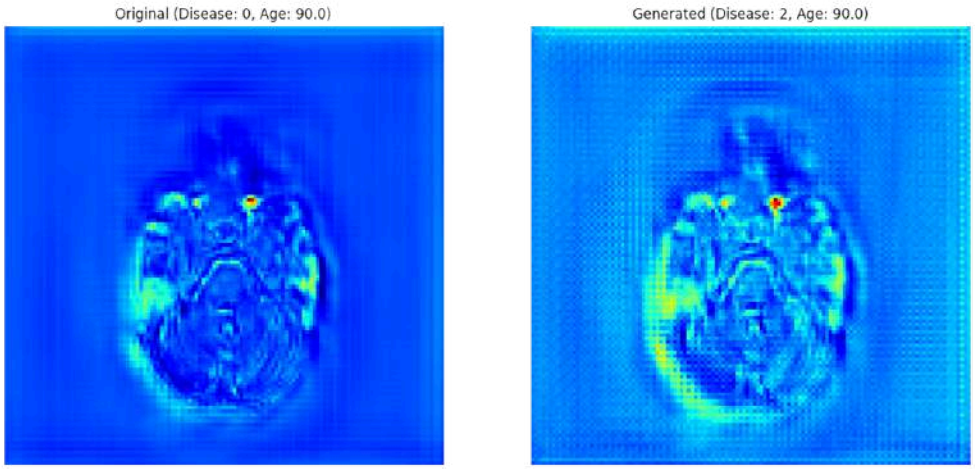


Figure 3.23: Counterfactual reconstruction based on disease label using the metadata-conditioned VAE. The original subject is 90 years old and cognitively normal. Instead of reconstructing the original SVF, the model is conditioned to generate a reconstruction for the same subject and age but with an Alzheimer's disease (AD) label. This illustrates the model's ability to answer counterfactual queries such as: "What would this patient's SVF look like if they had AD instead?"

Metric	**Value**
Full Latent Silhouette	-0.0140
Metadata Silhouette	-0.0206
Content Silhouette	-0.0109
Samples	1000

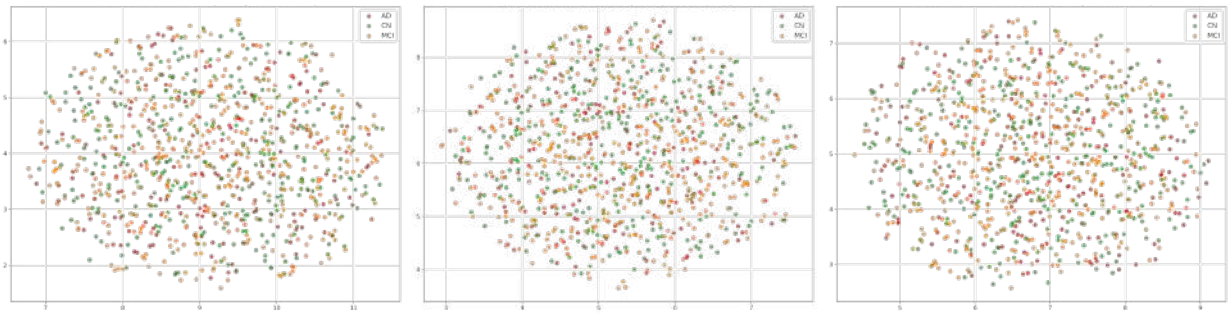
((a)) Clustering metrics across latent, metadata, and content spaces.

Group	**Count (%)**
AD	190 (19.0%)
CN	361 (36.1%)
MCI	449 (44.9%)

((b)) Group distribution in clustered latent space.

Table 3.18: Latent space clustering evaluation using silhouette scores and diagnostic group proportions.

These results suggest that disease-driven anatomical variation is either subtle or not sufficiently captured by the current embedding structure, even with contrastive supervision.



((a)) Full embedding latent space for 1k samples.

((b)) Only metadata embedding latent space for 1k samples.

((c)) Only content embedding latent space for 1k samples.

Figure 3.24: Visualization of the latent space components of the FactorizedMetadataVAE model trained with contrastive loss. Despite the contrastive loss, neither of the latent spaces shows separation based on disease label.

Split	Total Loss	Recon Loss	KLD Meta	KLD Content	Disease Loss	Age Loss	Indep. Loss	Contrastive Loss
Test	7.1449	0.2904	0.0569	0.0001	1.0575	57.6403	0.0139	1.5116
Validation	7.3595	0.1190	0.0482	0.0008	1.1280	61.6216	0.0137	1.5409

Table 3.19: Loss components of the metadata-conditioned VAE with contrastive learning on test and validation sets.

Split	Disease Prediction Accuracy
Test	0.4464
Validation	0.2431

Table 3.20: Evaluation metric (disease classification accuracy) for the metadata-conditioned VAE on test and validation sets.

3.6 Conclusion

In this chapter, we introduced **DIVA**—a Disentangled, metadata-Informed Variational Autoencoder framework designed for encoding stationary velocity fields (SVFs) from longitudinal brain MRI data. Extending the deformation-based generative paradigm established by SynBADD, DIVA incorporates clinically relevant metadata, such as age and diagnostic category, directly into the latent space. This enables the model to generate temporally consistent and clinically interpretable representations of neuroanatomical change.

We began by investigating the role of synthetic augmentation in learning robust representations. Our results showed that while augmentation improves reconstruction fidelity and promotes intra-subject consistency in vanilla VAEs, it alone does not induce meaningful temporal or clinical organization in the latent space. To address this, we progressively introduced increasingly structured models that incorporated metadata conditioning and latent factorization, culminating in architectures that explicitly disentangle anatomical content from temporal or diagnostic signals.

Applying this framework to temporal modeling, we demonstrated that factorized VAEs, combined with independence and contrastive objectives, yield latent spaces that are better aligned with chronological progression and subject identity. These structured embeddings supported smoother latent trajectories and improved interpretability. However, when extending the same approach to metadata conditioning—using age and diagnosis as supervisory signals—the results were more mixed. Although factorization facilitated modest gains in disease label separability and enabled counterfactual synthesis, contrastive training failed to produce clearly delineated clusters based on diagnosis. This suggests that disease-related deformation patterns are more heterogeneous and less spatially distinct than temporal trends, and that clinical labels alone may be insufficient to drive strong disentanglement in the deformation field space.

Nonetheless, DIVA offers a principled and extensible framework for learning semantically structured representations with minimal trade-off in reconstruction performance. Its ability to support counterfactual querying further underscores its potential for downstream clinical applications, though the reliability of such interpretations hinges on the strength of the latent disentanglement.

Finally, while metadata conditioning provides a foundation for clinical reasoning in the latent space, classifying subjects based on diagnostic labels remains inherently challenging. Distinguishing between conditions such as Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) likely requires modeling individualized disease trajectories that capture subtle, subject-specific patterns of neurodegeneration. In the following chapter, we address this challenge by leveraging VAE-derived latent embeddings for downstream classification, and propose a temporal modeling strategy tailored to subject-specific disease progression.

Chapter 4

Synthesize, then Analyze: Modeling Brain Aging and Disease Trajectories through Data Imputation

4.1 Introduction

Evaluating generative models solely through traditional quantitative metrics or qualitative visual inspection often presents an incomplete picture of model performance. Image-level metrics, while useful for benchmarking, tend to condense information into summary statistics and do not capture whether the synthesized outputs are clinically or semantically meaningful—particularly in domains like medical imaging where subtle structural changes over time are crucial. Conversely, qualitative assessments, though intuitive, are inherently subjective and may fail to detect fine-grained anatomical variations, especially in longitudinal studies where changes across timepoints can be subtle and easily overlooked.

To address these limitations, it is essential to move beyond conventional evaluation strategies and adopt frameworks that assess the utility of generative outputs in clinically or analytically relevant contexts. The *analysis-by-synthesis* paradigm offers such a mechanism: rather than merely judging outputs by their appearance or similarity to ground truth, it evaluates generative models based on their effectiveness in supporting downstream analytical tasks. This approach not only tests the realism of the generated data but also probes whether the model has learned meaningful and disentangled representations that align with domain knowledge and practical clinical objectives.

The *analysis-by-synthesis* paradigm is a framework for evaluating generative models by observing their behavior in downstream analytical tasks. Originally proposed in vision science and later embraced in graphics and machine learning [72, 73], this approach evaluates a model not solely by its ability to replicate data distributions, but by how its synthesized outputs enable or improve subsequent inference. In medical imaging, where interpretability and anatomical plausibility are essential, analysis-by-synthesis provides a principled route to assess the clinical relevance of learned generative representations [68, 70].

In this chapter, we use this framework to evaluate two complementary aspects of our work. First, we investigate the quality of anatomical synthesis produced by SynBADD—a deformation-based generative model for missing brain scan imputation. Second, we examine the semantic structure and clinical utility of the latent space learned by DIVA, our metadata-informed variational autoencoder.

We begin by evaluating SynBADD and baseline models along clinically grounded metrics. Specifically, we assess Dice overlap scores in eight neuroanatomical regions that are closely associated with the

onset and progression of Alzheimer's Disease: the hippocampus [74], parahippocampal gyrus, third ventricle [75], amygdala [76], entorhinal cortex [77], fusiform gyrus, inferior temporal gyrus, and middle temporal gyrus [78,79]. These regions serve as biomarkers for staging disease severity and predicting conversion from MCI to AD. While Dice scores reflect spatial agreement between predicted and ground-truth segmentations, they may not fully capture structural deviations. Hence, we also evaluate *volumetric fidelity*—quantifying regional brain volume in mm^3—to detect under- or overestimation of neurodegeneration.

Beyond individual scan quality, we use the extracted regional volumes to model subject-specific anatomical trajectories. By comparing the predicted trajectories from different models to the ground-truth evolution of atrophy, we assess whether the generative frameworks can capture meaningful structural change over time. These trajectory analyses offer a longitudinal perspective on model fidelity, aligning more directly with the real-world use case of tracking disease progression.

Having analyzed generation quality, we turn our attention to the structure of latent representations learned by DIVA. While conventional VAEs focus on pixel-wise reconstruction loss, they often entangle anatomical variation in a latent space that lacks semantic organization. In contrast, DIVA enforces structured representations through temporal and clinical conditioning, aiming to separate temporally evolving changes from static anatomical features.

To evaluate the semantic structure and clinical informativeness of these latent spaces, we adopt two downstream tasks. The first mirrors our earlier objective: synthesizing missing timepoints in longitudinal MRI sequences. We evaluate this task using the vanilla VAE variants—one trained on 1k augmented samples and another on a more densely augmented dataset of 16k synthetic deformations. These experiments qualitatively assess the expressive capacity of the learned latent space in guiding realistic anatomical synthesis, particularly in sparse data regimes.

The second task involves classifying disease stage—Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), or Cognitively Normal (CN)—from sequences of SVFs. Here, we leverage the vanilla VAE encoders to map longitudinal deformation sequences into latent trajectories. These latent embeddings are passed through a transformer with cross-attention pooling, followed by a classification layer. Remarkably, this pipeline achieves near-perfect performance, underscoring the discriminative power of SVF-based latent representations. More importantly, this approach offers practical advantages: it drastically reduces the dimensionality of input data, handles missing timepoints, accommodates class imbalance, and aligns closely with the individualized disease modeling needed in clinical neuroimaging [80, 81].

Taken together, these experiments demonstrate that deformation-based synthesis and structured latent modeling are not only effective generative strategies, but also clinically useful analytical tools. The analysis-by-synthesis lens helps ground these methods in real-world applicability—bridging the gap between algorithmic performance and meaningful clinical impact.

4.2 Implementation Details

All experiments were conducted using a fixed train/validation/test split, with subject-wise stratification to ensure no subject appeared in more than one subset. The training set consisted of 1,140 real samples from

275 subjects, the validation set included 121 samples from 30 subjects, and the held-out test set comprised 315 samples from 77 subjects.

The VAE model was trained independently using a batch size of 32. It was optimized using Adam with a learning rate of 5×10^{-5}, a ReduceLROnPlateau scheduler (patience = 5), and early stopping (patience = 10). Once trained, the VAE was frozen for downstream experiments.

For all synthesis tasks in the SynBADD pipeline, the full end-to-end model was trained with a batch size of 6, using the same optimizer and learning rate. Training incorporated a ReduceLROnPlateau scheduler with patience = 5 and early stopping with patience = 7.

All models were trained on a workstation with four NVIDIA RTX 6000 GPUs (48GB VRAM each). The VAE training set included 1,140 real SVFs and 15,760 synthetic augmentations. Augmentations included random SO(3) rotations (via Euler angles), affine ramp deformations, and elastic noise, all designed to preserve diffeomorphic structure.

Subjects contributed between 3 to 5 timepoints. Missing timepoints were masked during training and inference, allowing the model to operate on incomplete sequences and reflect real-world clinical scenarios.

4.3 Evaluation of Anatomical Synthesis in SynBADD via Region-Based Metrics

To assess the clinical and anatomical fidelity of synthesized scans, we evaluate SynBADD against NODER across a set of eight brain regions strongly associated with neurodegenerative progression in Alzheimer's disease: the hippocampus [82], parahippocampal gyrus [77], third ventricle [79], amygdala [83, 84], entorhinal cortex [77], fusiform gyrus [85], inferior temporal gyrus [78], and middle temporal gyrus [86, 87]. These regions were selected based on established associations with early atrophy and volumetric changes in AD progression.

4.3.1 Segmentation Accuracy via Dice Score.

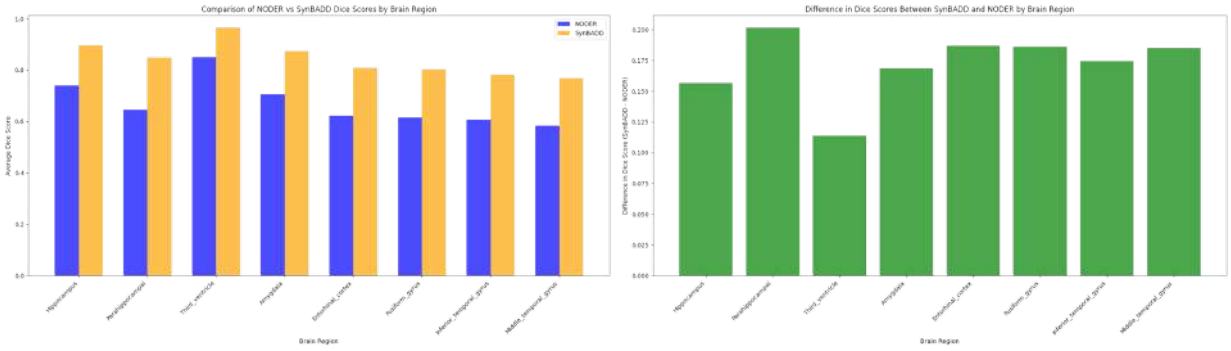
We first compute Dice similarity coefficients between predicted and ground-truth segmentations across the aforementioned regions. As summarized in Table 4.1, SynBADD consistently outperforms NODER with substantial improvements observed in medial temporal lobe structures, such as the parahippocampal gyrus and entorhinal cortex. The distribution of Dice scores is visualized in Figure 4.1(a), while the per-region improvements (SynBADD minus NODER) are shown in Figure 4.1(b), highlighting areas where SynBADD achieves the most significant anatomical fidelity gains.

4.3.2 Volumetric Accuracy.

While segmentation overlap provides one axis of evaluation, we additionally analyze how well each model preserves the volumetric properties of target regions. Table 4.2 reports both absolute and relative differences in estimated volumes (in mm^3) between predictions and ground truth. Notably, SynBADD reduces the error by orders of magnitude in several regions compared to NODER. For example, in the fusiform

Region	NODER Dice	SynBADD Dice	Dice Difference
Hippocampus	0.7397	**0.8965**	0.1567
Parahippocampal	0.6464	**0.8480**	0.2016
Third Ventricle	0.8499	**0.9635**	0.1136
Amygdala	0.7047	**0.8742**	0.1695
Entorhinal Cortex	0.6213	**0.8082**	0.1869
Fusiform Gyrus	0.6145	**0.8013**	0.1867
Inferior Temporal Gyrus	0.6066	**0.7816**	0.1750
Middle Temporal Gyrus	0.5819	**0.7670**	0.1851

Table 4.1: Dice scores for different brain regions comparing NODER and SynBADD.



((a)) Comparison of Dice scores between ground-truth ROIs and predictions from SynBADD (ours) and NODER.across regions of interest.

((b)) Difference in Dice scores (SynBADD minus NODER)

Figure 4.1: Quantitative comparison of segmentation accuracy using Dice similarity coefficient. (a) shows the absolute Dice scores for SynBADD and NODER methods across various brain regions. (b) presents the difference in Dice scores, illustrating the improvement achieved by SynBADD over NODER.

gyrus and inferior temporal gyrus—regions known to atrophy in early AD—SynBADD's predictions are within 2% of the true volume, whereas NODER underestimates by over 5-6%. Figure 4.2 illustrates these comparisons across all regions.

4.3.3 Longitudinal Trajectory Validation.

To evaluate temporal consistency and subject-specific modeling, we leverage the predicted segmentations to estimate volumetric trajectories across time. For each region and subject, we use ROI-based segmentation masks to compute region volumes at each predicted timepoint and compare them to ground-truth trajectories. SynBADD's predicted trajectories are plotted as solid lines, NODER's as faded lines, and the ground truth as dashed curves. Figures 4.3 through 4.10 visualize the results for all eight regions.

In nearly all regions and across multiple subjects, SynBADD tracks the true atrophy trajectory more accurately than NODER. For instance, in the hippocampus (Figure 4.3), SynBADD's trajectory maintains the

Region	NODER Diff (mm^3) ↓	NODER % Diff ↓	SynBADD Diff (mm^3) ↓	SynBADD % Diff ↓
Hippocampus	316.34	5.65%	**111.85**	**1.86%**
Parahippocampal	181.26	6.45%	**79.26**	**2.54%**
Third Ventricle	109.79	6.91%	**16.76**	**1.17%**
Amygdala	126.86	6.92%	**50.46**	**2.47%**
Entorhinal Cortex	178.01	7.83%	**91.02**	**3.85%**
Fusiform Gyrus	635.28	5.62%	**252.88**	**2.29%**
Inferior Temporal Gyrus	923.76	5.93%	**321.98**	**2.03%**
Middle Temporal Gyrus	1107.38	6.09%	**256.79**	**1.49%**

Table 4.2: Absolute and relative differences in volume prediction between NODER and SynBADD methods across regions. Lower values indicate better performance. Bold values show the lower (better) difference in each row.

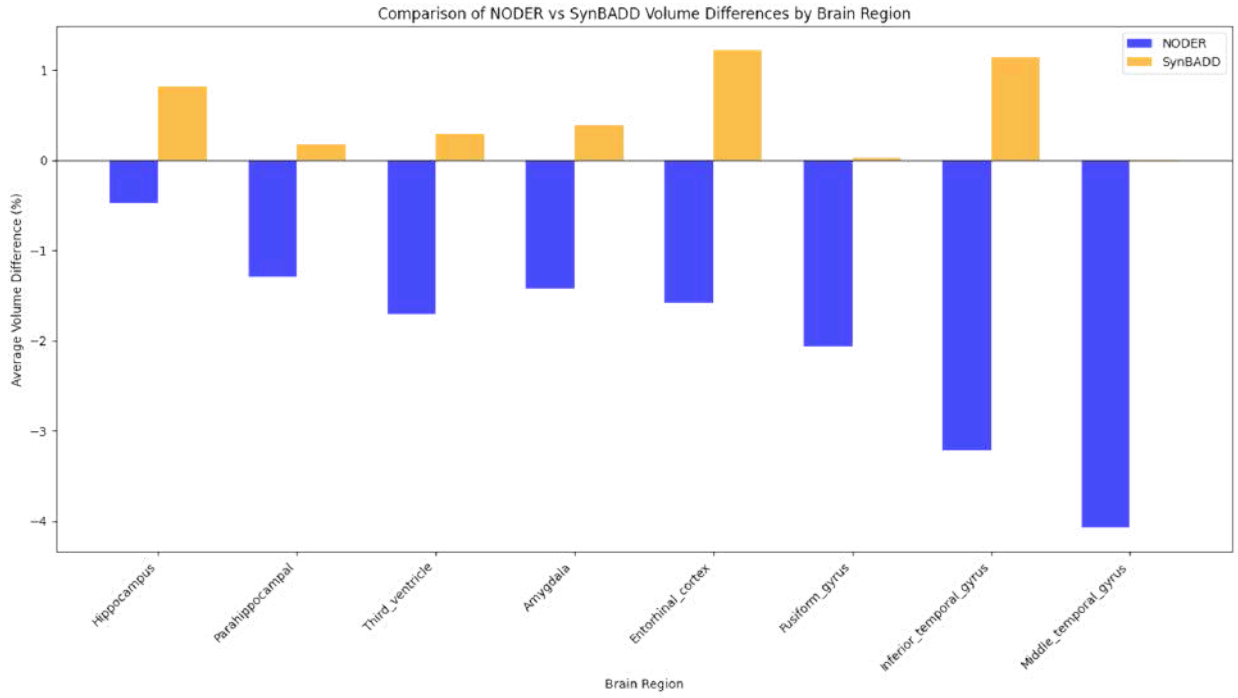


Figure 4.2: Comparison of absolute differences in estimated brain region volumes (in mm^3) between predicted segmentations and ground-truth ROIs for SynBADD (ours) and NODER. This analysis highlights the accuracy of volumetric estimation across different anatomical structures.

correct slope and intercept over time, while NODER tends to underpredict atrophy and exhibits inconsistent trends. These results further validate SynBADD's capacity to model continuous anatomical change by synthesizing plausible intermediate states between sparse real observations.

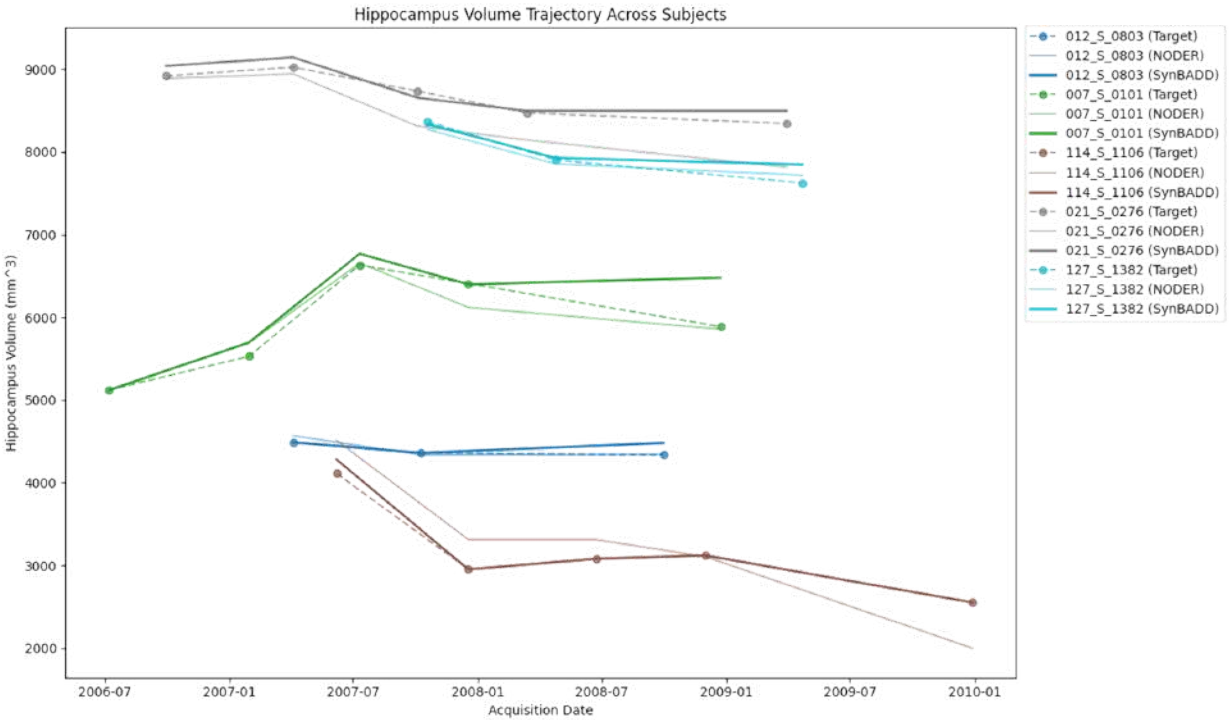


Figure 4.3: Hippocampus Volume trajectory for 5 subjects.

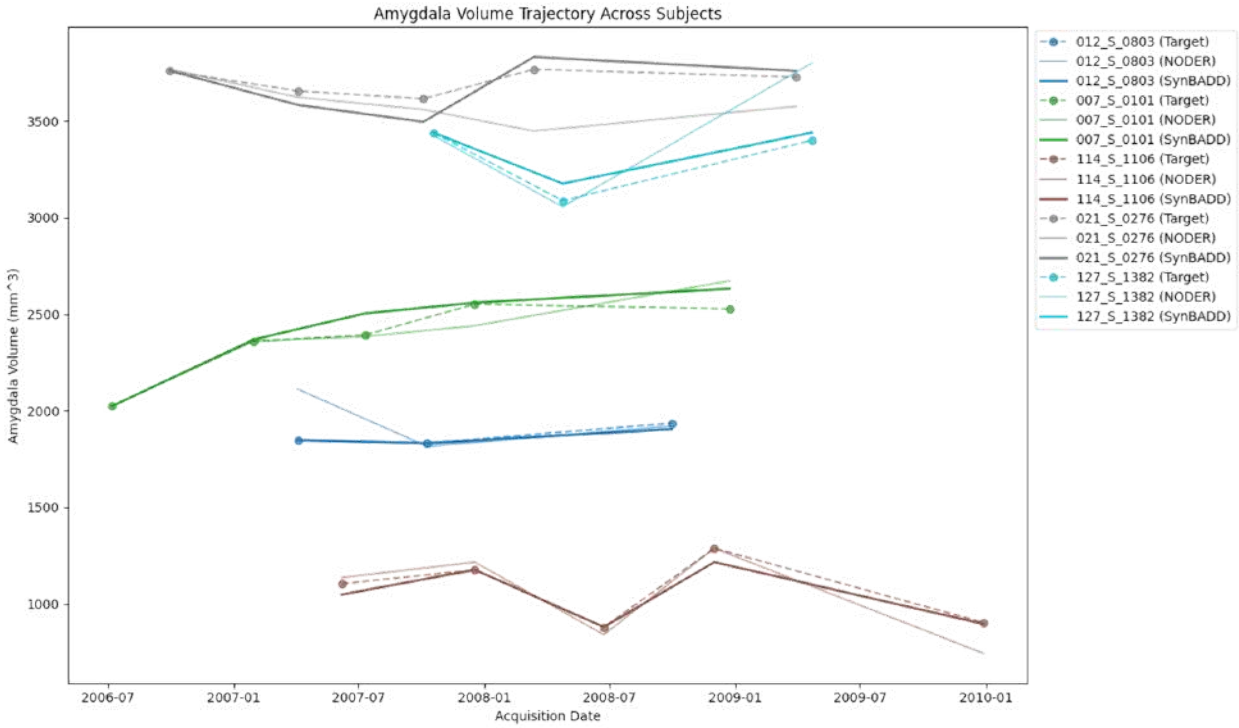


Figure 4.4: Amygdala Volume trajectory for 5 subjects.

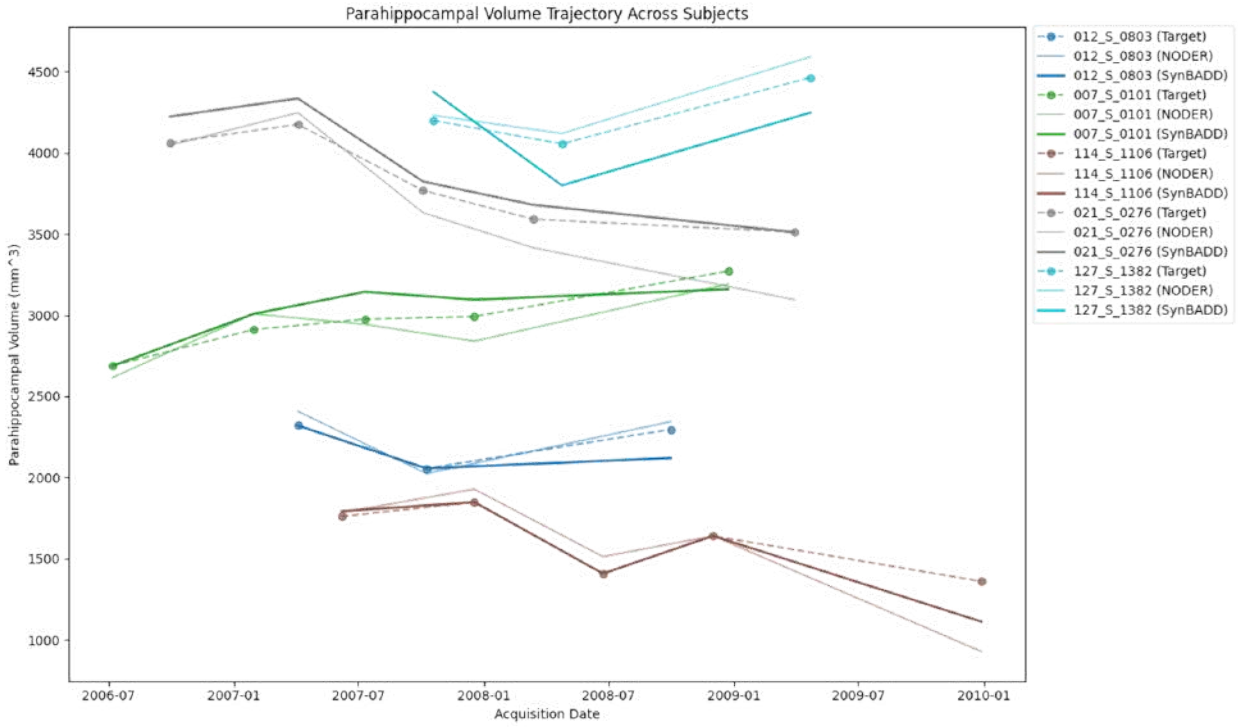


Figure 4.5: Parahippocampal gyrus Volume trajectory for 5 subjects.

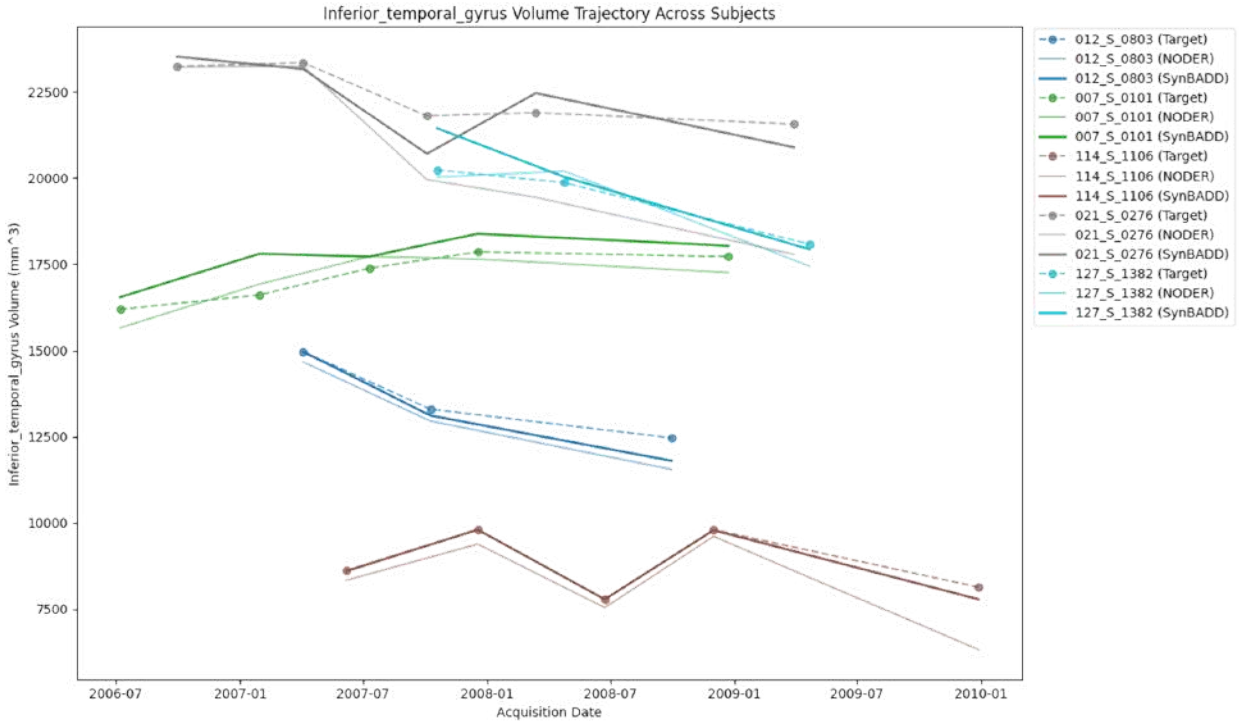


Figure 4.6: Inferior temporal gyrus Volume trajectory for 5 subjects.

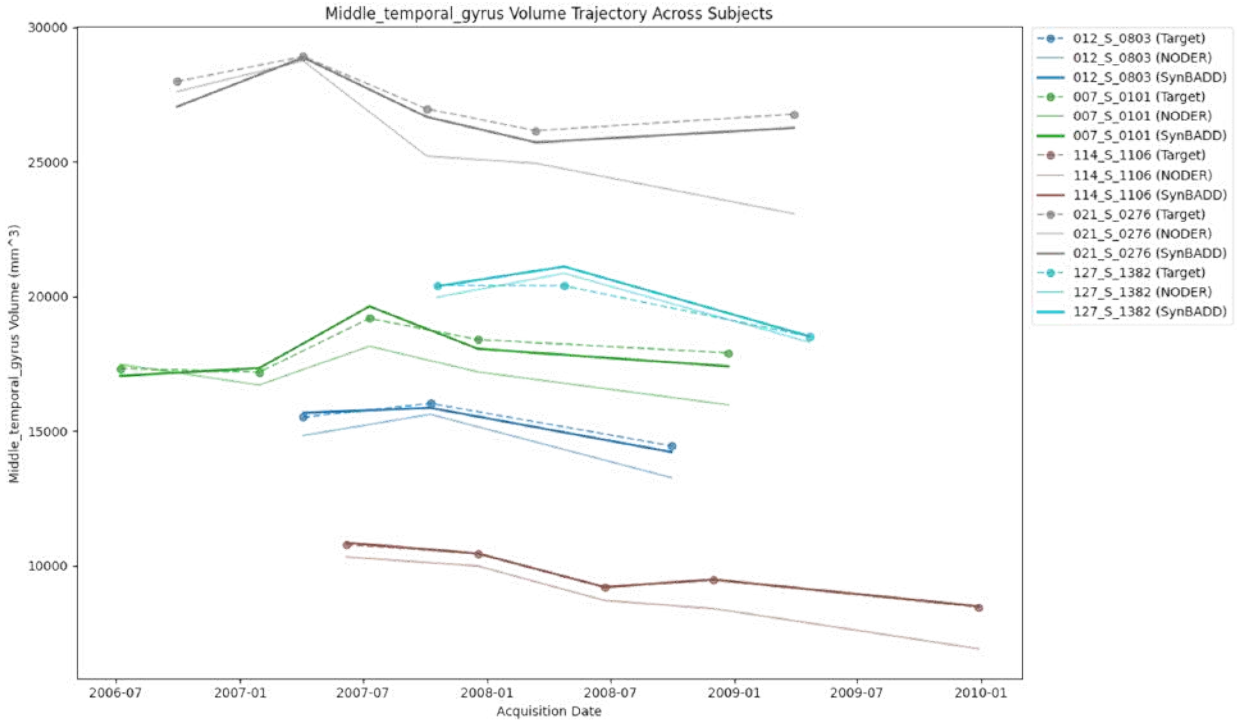


Figure 4.7: Middle temporal gyrus Volume trajectory for 5 subjects.

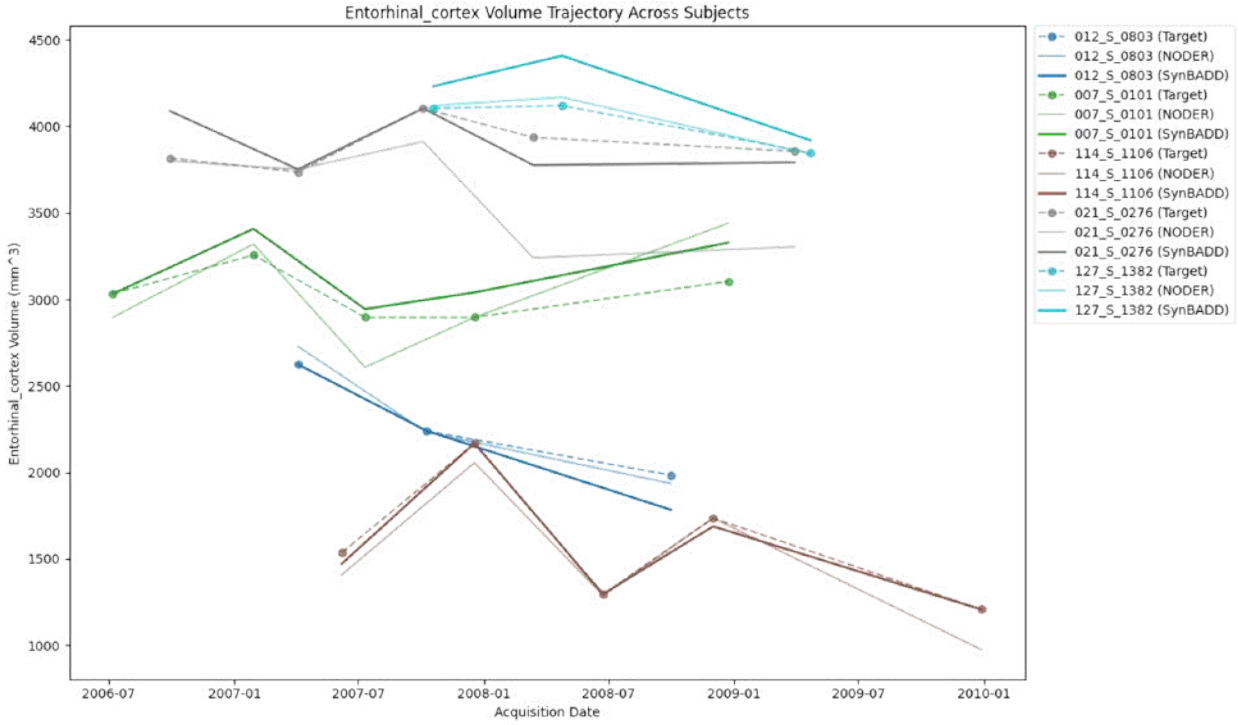


Figure 4.8: Entorhinal Cortex Volume trajectory for 5 subjects.

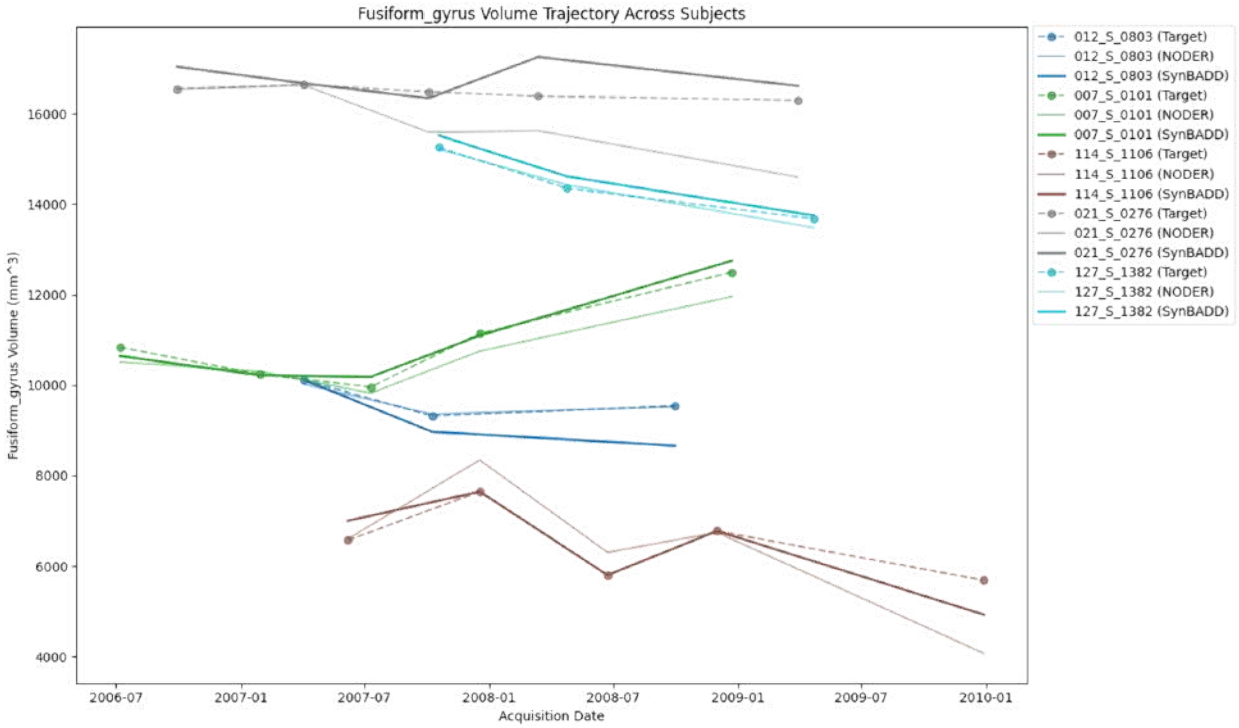


Figure 4.9: Fusiform gyrus Volume trajectory for 5 subjects.

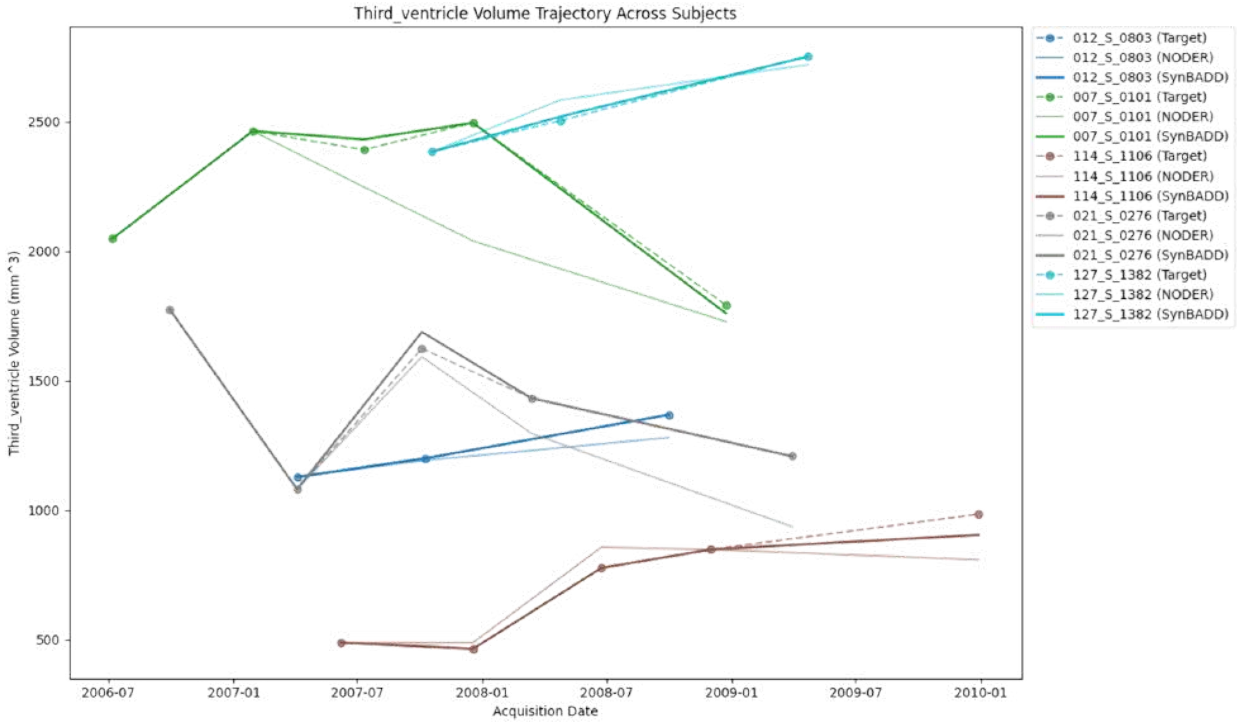


Figure 4.10: Third Ventricle Volume trajectory for 5 subjects.

4.3.4 Image-Level Quality Metrics.

Finally, to contextualize these region-specific evaluations within broader image synthesis quality, we report PSNR and SSIM scores in Table 4.3. SynBADD exhibits higher fidelity reconstructions across all metrics, further reinforcing that anatomically informed, deformation-based synthesis produces outputs that are both visually realistic and structurally accurate.

Together, these analyses demonstrate that SynBADD not only enhances anatomical segmentation fidelity but also captures the dynamics of longitudinal progression with greater accuracy than previous methods. This enables a more clinically meaningful modeling of subject-specific disease trajectories, even when intermediate scans are missing or corrupted.

Metric ↑	**NODER**	**SynBADD**	**Improvement (%)**
PSNR	30.31	33.32	9.89%
SSIM	0.967	0.9864	2%
Dice Hippocampus	0.7397	0.8962	21.12%
Dice Parahippocampal	0.6463	0.8476	31.20%
Dice Third Ventricle	0.8493	0.9633	13.41%
Dice Amygdala	0.7049	0.8735	23.89%
Dice Entorhinal Cortex	0.6215	0.8082	30.06%
Dice Fusiform Gyrus	0.6150	0.8010	30.24%
Dice Inferior Temporal Gyrus	0.6070	0.7815	28.74%
Dice Middle Temporal Gyrus	0.5820	0.7672	31.80%

Table 4.3: Percentage improvement of SynBADD over NODER across PSNR, SSIM, and Dice scores for various brain regions.

4.4 Latent Evaluation via Downstream Tasks in DIVA

While anatomical synthesis provides important validation of spatial and volumetric fidelity, it does not directly measure how well the latent space captures disease-relevant representations. To evaluate the clinical utility of our VAE-derived latent encodings, we assess DIVA's performance on two complementary downstream tasks: high-resolution generation and disease classification.

4.4.1 Fine-Grained Structural Fidelity.

First, we assess DIVA's reconstruction quality via SynBADD, to evaluate whether it is suitable enough to synthesize subtle yet clinically meaningful anatomical differences when predicting missing scans. DIVA acts as a backbone for encoding the longitudinal SVF sequence for each subject. SynBADD takes this encoded sequence as input and predicts the latent for the missing timepoint as the output. DIVA then decodes

this latent to reconstruct the full-resolution SVF which after refinement is used to warp the baseline scan. Figure 4.11 illustrates qualitative results for a subject scan synthesized by warping the baseline using the predicted deformation field. Green circles mark regions where the warped image faithfully captures fine-grained atrophic progression, closely aligning with the ground-truth target. Red circles highlight regions in the baseline scan that differ from the warped output, emphasizing how the deformation successfully accounts for structural changes. This shows that the latent representations encode not only coarse spatial deformations, but also reflect subtle temporal changes that are diagnostically informative.

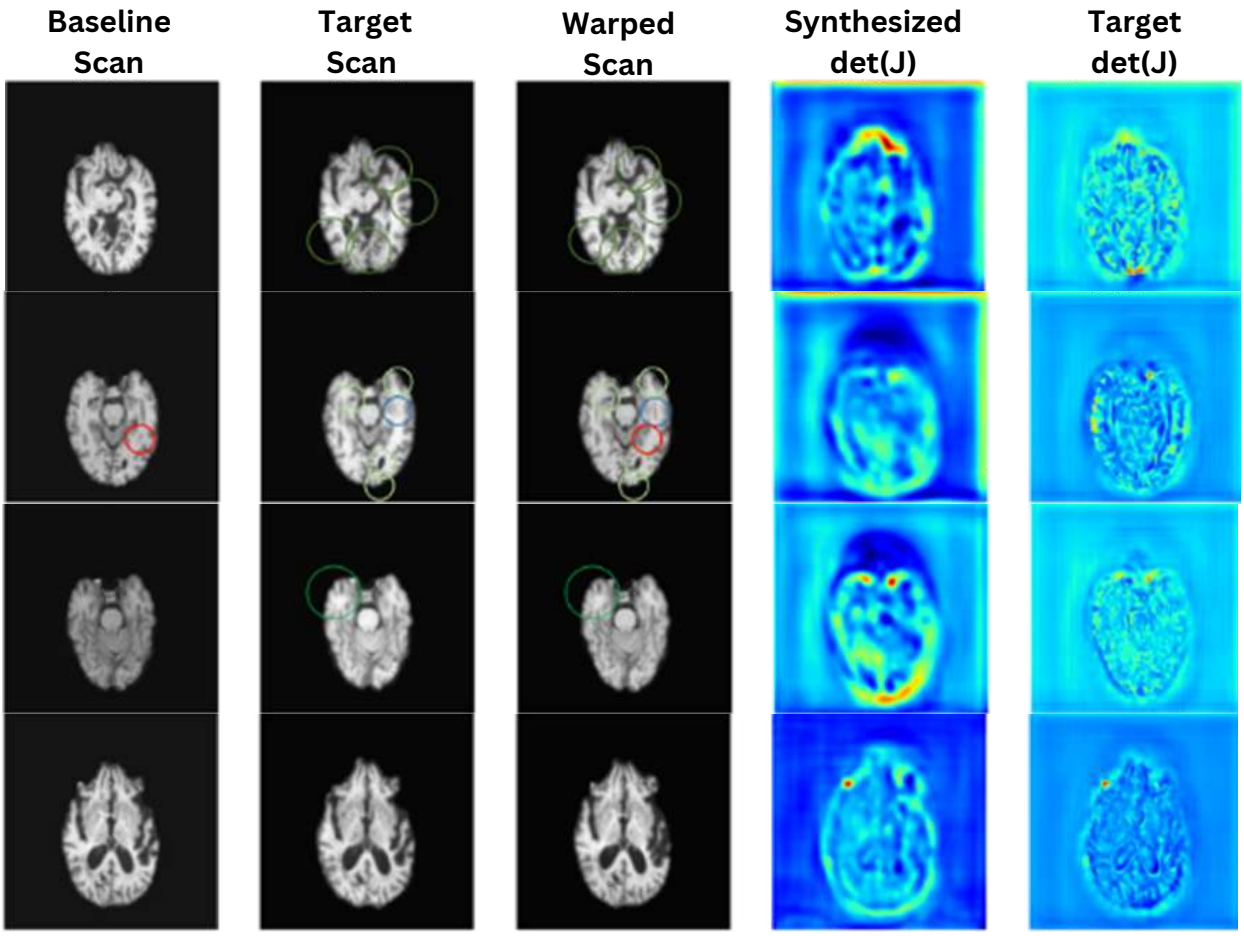


Figure 4.11: Qualitative illustration of SynBADD's capability to capture subtle anatomical changes. Green circles highlight regions where fine-grained atrophy is accurately captured when warping the baseline scan to the target scan. Red circles indicate the corresponding anatomical differences observed between the warped scan and the original baseline scan, emphasizing the alignment quality and sensitivity to structural changes.

4.4.2 Disease Classification from Latent Trajectories.

We further evaluate the latent space through a challenging three-way classification task distinguishing Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) subjects. Each subject's full longitudinal sequence of SVFs (one per timepoint) is encoded by the trained VAE into a

compact latent representation. A transformer with cross-attention pooling then aggregates these timepoint-wise embeddings into a subject-level representation, which is passed through a classifier to predict diagnostic category.

This classification task serves as a proxy for how well the latent space organizes disease progression patterns. Notably, the model is trained without access to diagnostic labels during VAE training—only using age and timepoint as metadata—which further validates the semantic alignment of the latent space with underlying disease dynamics. As reported in Table 4.4, the model achieves exceptionally high classification accuracy, correctly classifying all test subjects in folds 1, 3, and 5, and making just a single misclassification in folds 2 and 4. This consistent performance demonstrates the model's robustness to missing timepoints and class imbalance, and its capacity to learn discriminative trajectories even in a highly compressed space.

By leveraging subject-specific deformation representations instead of raw image intensities, the model gains interpretability and anatomical coherence. These compact trajectories enable more effective modeling of neurodegenerative patterns and align closely with clinical workflows in Alzheimer's disease staging [63, 65, 88], reinforcing the value of deformation-based latent modeling as both a generative and analytic tool.

Fold	Number of Test Subjects	Accuracy	Number of Misclassifications
1	77	100.00%	0
2	77	98.72%	1
3	77	100.00%	0
4	77	98.72%	1
5	76	100.00%	0
Average Accuracy	–	**99.49%**	–
Average Misclassifications	–	–	**0.0052**

Table 4.4: Downstream classification performance to evaluate the representational quality of the VAE-encoded features. We perform a three-way classification task to distinguish Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) subjects based on subject-specific disease trajectories estimated from SVF scans. The model demonstrates strong generalization across folds, effectively handling missing timepoints, class imbalance, and inter-subject variability, resulting in consistently high classification accuracy.

4.4.3 Are These Models Clinically Robust and Practically Usable?

4.4.3.1 Robustness to Sparse and Incomplete Data

Both SynBADD and DIVA operate reliably on incomplete sequences. Their performance remains stable even in the absence of certain timepoints and variable length of input sequence.

4.4.3.2 Generalization Across Subjects and Imbalanced Classes

DIVA performs well across subject groups and diagnostic classes, despite class imbalance (Refer Table 4.5). This indicates the model's ability for generalization and robustness.

Group	Count (%)
AD	190 (19.0%)
CN	361 (36.1%)
MCI	449 (44.9%)

Table 4.5: Group distribution in clustered latent space.

4.4.3.3 Integration Potential in Clinical Workflows

SynBADD produces interpretable outputs since we make use of the subject's baseline scan and warp it with deformation fields. DIVA provides a method to encode compact latent embeddings—that can be used in clinical settings for tracking, visualization, or automated diagnosis as shown through the various downstream tasks.

4.5 Conclusion and Discussion

This chapter demonstrated how anatomically grounded generative modeling and structured latent encoding can jointly advance the analysis of longitudinal brain MRI data. Through the SynBADD framework, we showed that operating in the stationary velocity field (SVF) space enables deformation-based synthesis of missing scans that is not only computationally efficient, but also anatomically faithful. SynBADD consistently outperformed baseline methods across multiple clinical metrics—achieving superior Dice scores in regions implicated in neurodegeneration, lower volumetric deviation, and improved temporal alignment of subject-specific trajectories.

Complementing this, the DIVA framework highlighted the representational strength of temporally and metadata-aware latent spaces. By structuring the latent embedding space to disentangle temporal and demographic variation, DIVA facilitated robust downstream classification of Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) subjects. The ability to model longitudinal patterns from high-dimensional SVFs while accommodating missing data reflects the flexibility and clinical utility of this approach.

Together, these results suggest that analysis-by-synthesis frameworks—when carefully grounded in anatomy and enriched with domain-specific supervision—can bridge the gap between generative modeling and clinical interpretation. They also underscore the broader promise of deformation-based generative paradigms in supporting both reconstruction and diagnostic reasoning within longitudinal neuroimaging pipelines.

Chapter 5

Conclusion

5.1 Conclusion

This thesis begins with a simple question: how can we meaningfully model anatomical change in the human brain, especially when that change is subtle, irregular, and only partially observed? Through the development of SynBADD and DIVA, we have shown that a deformation-based, structured generative framework—guided by clinical knowledge and temporal awareness—can offer both scientific insight and practical utility.

We explored this question by shifting the modeling focus from raw image intensities to deformation fields—specifically, stationary velocity fields (SVFs)—which encode neuroanatomical changes as smooth, structure-preserving transformations. This formulation not only respects biological plausibility but also enables more data-efficient and interpretable modeling. Rather than memorizing appearances, our models learn how anatomy evolves, capturing meaningful patterns of brain atrophy across time.

The proposed methods—SynBADD for deformation-based synthesis and DIVA for metadata-informed latent encoding—work together to address key challenges in longitudinal neuroimaging: missing data, subject variability, and clinical interpretability. SynBADD generates anatomically faithful intermediate scans from sparse timepoints, while DIVA learns temporally and demographically structured embeddings that support downstream tasks like disease classification and counterfactual synthesis.

5.1.1 Key Findings and Insights

Our experiments revealed several important insights about modeling anatomical progression using deformation-based generative frameworks:

- **Deformation over intensity:** Representing brain changes as SVFs, rather than directly modeling image intensities, leads to more anatomically consistent and temporally coherent synthesis. This shift provides robustness to intensity variability and enhances interpretability.
- **Value of augmentation:** Synthetic deformation augmentations improved intra-subject consistency and generalization in vanilla VAEs. However, they were insufficient for inducing clinically meaningful latent structure, highlighting the need for stronger supervision.

- **Disentanglement via metadata:** Conditioning latent spaces on age, diagnosis, and timepoint enabled partial disentanglement of anatomical variation. Factorized representations helped isolate temporal or clinical trends from static subject-specific anatomy, supporting tasks like counterfactual generation and trajectory modeling.

- **Limitations of contrastive learning:** While contrastive losses enhanced temporal alignment in time-conditioned VAEs, they were less effective in the metadata setting. Disease labels like AD or MCI may not manifest as distinct deformation patterns in SVF space, suggesting the need for more expressive supervision or richer representations.

- **Clinical utility:** The learned latent spaces, particularly from structured models like DIVA, supported disease classification and plausible counterfactual inference. These capabilities demonstrate the potential of deformation-based generative models as both analytical and assistive tools in neuroimaging.

Together, these findings show that we can obtain semantically meaningful representations with disentangled latent modeling which provides a viable framework for learning from sparse, heterogeneous clinical data.

5.1.2 Limitations

While the models presented in this thesis demonstrate promising results, several limitations remain:

- **Limited expressiveness of clinical supervision:** Metadata such as diagnostic labels (CN, MCI, AD) were used to guide the latent space, but these labels are inherently coarse and may not reflect the nuanced and continuous spectrum of neurodegenerative progression. As a result, the learned latent spaces exhibited only limited diagnostic separability. Future iterations of DIVA could incorporate richer forms of supervision, such as full longitudinal trajectories across timepoints, as successfully employed in our disease classification experiments.

- **Imaging modality:** All experiments were conducted using T1-weighted structural MRI. While suitable for capturing brain atrophy, T1-weighted imaging is not the clinical standard for diagnosing Alzheimer's Disease or related conditions. In contrast, positron emission tomography (PET)—particularly tau- and amyloid-PET—provides direct measurements of disease pathology. Extending the proposed framework to model such modalities would enhance biological interpretability and clinical relevance.

- **Metadata scope:** The current implementation conditions only on age and diagnostic category. However, disease progression is multifactorial, influenced by genetic risk (e.g., APOE4 allele presence), pharmacological interventions, and cognitive assessments. Integrating such metadata—available in extended subsets of the ADNI dataset—could enable more personalized modeling and improve the specificity of learned representations.

- **Dataset diversity:** All models were trained and validated on a single dataset focused on Alzheimer's Disease. While results are encouraging, their generalizability to other neurodegenerative diseases, age

groups, or clinical settings remains untested. Future work should evaluate performance across diverse datasets to assess robustness and transferability.

- **Counterfactual synthesis:** Although the proposed framework supports counterfactual generation via metadata manipulation, there is no established ground truth for validating such outputs. The realism and clinical plausibility of counterfactuals remain difficult to assess. A potential avenue is to evaluate generated outputs using independent predictors—for example, existing models trained to estimate age or classify diagnosis from anatomy. Consistency between intended metadata and predictions could serve as a proxy for evaluating counterfactual quality.

Addressing these limitations will require both methodological advancements and access to richer, longitudinal datasets with broader clinical, genetic, and biomarker annotations.

5.1.3 Future Work

Building upon the contributions of this thesis, several promising directions can be pursued to enhance the clinical applicability and generalizability of the proposed framework:

- **Integration of biologically informative modalities:** Future work should extend DIVA and SynBADD to operate on PET imaging, particularly tau- and amyloid-PET, which offer direct insights into Alzheimer's pathology. Modeling deformation trajectories in these modalities could lead to more biologically grounded representations of disease progression.

- **Multimodal fusion:** A natural extension involves integrating structural MRI with PET, genetic, and cognitive data to learn joint latent spaces. Such fusion could improve model robustness and enable comprehensive disease modeling by capturing complementary aspects of neurodegeneration.

- **Personalized modeling through enriched metadata:** Incorporating additional metadata—such as APOE genotype, medication history, and cognitive scores—would allow for more individualized modeling of brain aging trajectories. This could support precision medicine approaches by adapting generative trajectories to subject-specific risk profiles.

- **Cross-dataset generalization:** To ensure clinical robustness, future research should validate models across heterogeneous datasets encompassing different scanners, demographics, and clinical cohorts. Domain adaptation techniques may be necessary to bridge distributional shifts and ensure model transferability.

- **Counterfactual verification and calibration:** While this work introduced counterfactual generation capabilities, verifying their clinical utility remains an open problem. Future efforts could combine counterfactual synthesis with external evaluators—such as clinicians or pretrained classifiers—to assess plausibility, uncertainty, and impact on downstream decision-making.

- **Trajectory-aware diagnosis models:** Finally, downstream classifiers could be further refined by explicitly incorporating temporal dynamics. Sequence models operating on latent trajectories, such

as recurrent or attention-based architectures, may improve diagnostic performance and support early detection of disease transitions.

These directions reflect a broader ambition: to move from static snapshots to temporally aware, clinically grounded representations of neurodegeneration—laying the foundation for interpretable, personalized, and trustworthy AI systems in healthcare.

Bibliography

[1] Alzheimer's Disease International, "Dementia statistics," https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/, 2025, accessed 12 April 2025.

[2] World Health Organization, "Dementia," https://www.who.int/news-room/facts-in-pictures/detail/dementia, 2023, accessed 31 July 2023.

[3] ——, "Dementia," https://www.who.int/news-room/fact-sheets/detail/dementia, 2025, accessed 12 April 2025.

[4] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S. Costafreda, A. Dias, N. Fox, L. Gitlin, R. Howard, H. Kales, E. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E. Sampson, Q. Samus, L. Schneider, G. Selbæk, L. Teri, and N. Mukadam, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *Lancet*, vol. 396, no. 10248, pp. 413–446, 2020. [Online]. Available: https://www.thelancet.com/article/S0140-6736(20)30367-6/fulltext

[5] E. McDade, G. Wang, J. Hassenstab, V. L. Buckles, A. M. Fagan, D. M. Holtzman, J. C. Morris, K. Paumier, R. J. Bateman, B. A. Gordon, T. L. S. Benzinger, D. S. Marcus, N. J. Cairns, A. M. Jefferson, R. Allegri, S. B. Berman, W. Klunk, J. Noble, J. Ringman, B. Ghetti, M. R. Farlow, R. A. Sperling, J. Chhatwal, S. Salloway, N. R. Graff-Radford, P. R. Schofield, C. L. Masters, M. N. Rossor, N. C. Fox, J. Levin, and M. Jucker, "Longitudinal cognitive and biomarker changes in dominantly inherited alzheimer disease," *Neurology*, vol. 91, no. 14, pp. e1295–e1306, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6177272/

[6] M. Gabel, R. M. Bollinger, D. W. Coble, J. D. Grill, D. F. Edwards, J. H. Lingler, E. Chin, and S. L. Stark, "Retaining participants in longitudinal studies of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 87, no. 2, pp. 945–955, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9673904/

[7] L. Zhou, F. Ma, Y. Liang, X. Hu, and D. Zhi, "An empirical evaluation of the impacts of missing ehr data in comparative effectiveness research," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10280351/, 2023, accessed 20 April 2025.

[8] B. A. Bell, S. G. West, I. McDowell, M. Doucet, J. S. de Bruin, and F. R. Hopp, "Missing data approaches in longitudinal studies of aging: A case example," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10249888/, 2023, accessed 20 April 2025.

[9] R. J. Little and D. B. Rubin, "Missing data: A special challenge in aging research," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2695652/, 2009, accessed 20 April 2025.

[10] S. Ghosal, S. Liu, D. Subramanian, T. Wang, K. Jung, W. Xu, C. Hegde, and J. Ghosh, "Explainability of machine learning models under missing data," https://arxiv.org/html/2407.00411v1, 2024, accessed 20 April 2025.

[11] Y. Skandarani, A. Alansary, S. Albarqouni, B. Glocker, D. Rueckert, T. Doel, K. Hammernik, and E. Konukoglu, "Gans for medical image synthesis: An empirical study," https://arxiv.org/abs/2105.05318, 2021, accessed 20 April 2025.

[12] Y. Li, K. Song, T. Qian, D. Chen, W. Xie, and T. Liu, "Zero-shot medical image translation via frequency-guided diffusion models," https://arxiv.org/abs/2304.02742, 2023, accessed 20 April 2025.

[13] K. Kadry, T. Goodwin, N. Hachach-Haram, E. Marban, K. Khalil, A. Esteva, and P. Rajpurkar, "Probing the limits and capabilities of diffusion models for the anatomic editing of digital twins," https://www.nature.com/articles/s41746-024-01332-0, 2024, accessed 20 April 2025.

[14] J. S. Yoon, C. Zhang, H.-I. Suk, J. Guo, and X. Li, "Sadm: Sequence-aware diffusion model for longitudinal medical image generation," in *International Conference on Information Processing in Medical Imaging*. Springer, 2023, pp. 388–400.

[15] C. Yuan, J. Duan, K. Xu, N. J. Tustison, R. A. Hubbard, and K. A. Linn, "Remind: recovery of missing neuroimaging using diffusion models with application to alzheimer's disease," *Imaging Neuroscience*, vol. 2, pp. 1–14, 2024.

[16] Y. Wu, T. Z. Jiahao, J. Wang, P. A. Yushkevich, M. A. Hsieh, and J. C. Gee, "Nodeo: A neural ordinary differential equation based optimization framework for deformable image registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 804–20 813.

[17] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[18] H. Bai and Y. Hong, "Noder: Image sequence regression based on neural ordinary differential equations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 673–682.

[19] L. Sun, J. Chen, Y. Xu, M. Gong, K. Yu, and K. Batmanghelich, "Hierarchical amortized gan for 3d high resolution medical image synthesis," *IEEE journal of biomedical and health informatics*, vol. 26, no. 8, pp. 3966–3975, 2022.

[20] G. Kwon, C. Han, and D.-s. Kim, "Generation of 3d brain mri using auto-encoding generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 118–126.

[21] A. B. Qasim, I. Ezhov, S. Shit, O. Schoppe, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Lipkova, H. Li, and B. Menze, "Red-gan: Attacking class imbalance via conditioned generation. yet another medical imaging perspective." in *Medical imaging with deep learning*. PMLR, 2020, pp. 655–668.

[22] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, "Conditional diffusion models for semantic 3d medical image synthesis," *imaging*, vol. 19, p. 20, 2023.

[23] M. Zhou, M. W. Wagner, U. Tabori, C. Hawkins, B. B. Ertl-Wagner, and F. Khalvati, "Generating 3d brain tumor regions in mri using vector-quantization generative adversarial networks," *Computers in Biology and Medicine*, vol. 185, p. 109502, 2025.

[24] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 10, p. e0140381, 2015.

[25] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, "Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and IMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1*. Springer, 2018, pp. 106–114.

[26] K. Han, Y. Xiong, C. You, P. Khosravi, S. Sun, X. Yan, J. S. Duncan, and X. Xie, "Medgen3d: A deep generative framework for paired 3d image and mask generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 759–769.

[27] P. Liu, O. Puonti, X. Hu, D. C. Alexander, and J. E. Iglesias, "Brain-id: Learning contrast-agnostic anatomical representations for brain imaging," in *European Conference on Computer Vision*. Springer, 2024, pp. 322–340.

[28] B. Huang, X. Wang, B. Jiang, L. Kong, H. Hou, and J. Zhou, "Case report: A neurolinguistic and neuroimaging study on a chinese follow-up case with logopenic-variant of primary progressive aphasia," *Frontiers in Neurology*, vol. 13, p. 963970, 2022.

[29] S. Fujita, S. Mori, K. Onda, S. Hanaoka, Y. Nomura, T. Nakao, T. Yoshikawa, H. Takao, N. Hayashi, and O. Abe, "Characterization of brain volume changes in aging individuals with normal cognition using serial magnetic resonance imaging," *JAMA Network Open*, vol. 6, no. 6, pp. e2 318 153–e2 318 153, 06 2023. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2023.18153

[30] M. Woodward, D. A. Bennett, T. Rundek, G. Perry, and T. Rudka, "The relationship between hippocampal changes in healthy aging and alzheimer's disease: a systematic literature review," *Frontiers in aging neuroscience*, vol. 16, p. 1390574, 2024.

[31] C. McDonald, L. McEvoy, L. Gharapetian, C. Fennema-Notestine, D. Hagler Jr, D. Holland, A. Koyama, J. Brewer, and A. Dale, "Regional rates of neocortical atrophy from normal aging to early alzheimer disease," *Neurology*, vol. 73, no. 6, pp. 457–465, 2009.

[32] K. K. Leung, J. W. Bartlett, J. Barnes, E. N. Manning, S. Ourselin, N. C. Fox, and A. D. N. Initiative, "Cerebral atrophy in mild cognitive impairment and alzheimer disease: rates and acceleration," *Neurology*, vol. 80, no. 7, pp. 648–654, 2013.

[33] H. Tabatabaei-Jafari, M. E. Shaw, and N. Cherbuin, "Cerebral atrophy in mild cognitive impairment: a systematic review with meta-analysis," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 4, pp. 487–504, 2015.

[34] Y. Uchida, K. Nishimaki, A. Soldan, A. Moghekar, M. Albert, and K. Oishi, "Acceleration of brain atrophy and progression from normal cognition to mild cognitive impairment," *JAMA Network Open*, vol. 7, no. 10, pp. e2 441 505–e2 441 505, 2024.

[35] R. Sivera, H. Delingette, M. Lorenzi, X. Pennec, N. Ayache, A. D. N. Initiative *et al.*, "A model of brain morphological changes related to aging and alzheimer's disease from cross-sectional assessments," *NeuroImage*, vol. 198, pp. 255–270, 2019.

[36] J. Sluimer, H. Vrenken, M. Blankenstein, N. Fox, P. Scheltens, F. Barkhof, and W. Van Der Flier, "Whole-brain atrophy rate in alzheimer disease: identifying fast progressors," *Neurology*, vol. 70, no. 19_part_2, pp. 1836–1841, 2008.

[37] P. Vemuri and C. R. Jack, "Role of structural mri in alzheimer's disease," *Alzheimer's research & therapy*, vol. 2, pp. 1–10, 2010.

[38] G. Hwang, A. Abdulkadir, G. Erus, M. Habes, R. Pomponio, H. Shou, J. Doshi, E. Mamourian, T. Rashid, M. Bilgel *et al.*, "Disentangling alzheimer's disease neurodegeneration from typical brain ageing using machine learning," *Brain Communications*, vol. 4, no. 3, p. fcac117, 2022.

[39] M. D. Silva, C. H. Sudre, K. Garcia, C. Bass, M. J. Cardoso, and E. C. Robinson, "Distinguishing healthy ageing from dementia: A biomechanical simulation of brain atrophy using deep networks," in *Machine Learning in Clinical Neuroimaging: 4th International Workshop, MLCN 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4*. Springer, 2021, pp. 13–22.

[40] A. M. Fjell and K. B. Walhovd, "Structural brain changes in aging: courses, causes and cognitive consequences," *Reviews in the Neurosciences*, vol. 21, no. 3, pp. 187–221, 2010.

[41] N. Raz *et al.*, "Regional brain changes in aging healthy adults: general trends, individual differences and modifiers," *Cerebral Cortex*, vol. 15, no. 11, pp. 1676–1689, 2005.

[42] L. K. McEvoy, C. Fennema-Notestine, J. C. Roddey *et al.*, "Structural neuroimaging for early diagnosis of alzheimer's disease," *Alzheimer's & Dementia*, vol. 5, no. 4, pp. 340–346, 2009.

[43] C. R. Jack *et al.*, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[44] J. Barnes *et al.*, "Atrophy rates of the hippocampus and whole brain in subjects with mild cognitive impairment with and without conversion to alzheimer's disease," *Neurobiology of Aging*, vol. 30, no. 4, pp. 674–681, 2009.

[45] B. C. Dickerson *et al.*, "The cortical signature of alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild ad dementia and is detectable in asymptomatic amyloid-positive individuals," *Cerebral Cortex*, vol. 19, no. 3, pp. 497–510, 2009.

[46] S. M. Nestor *et al.*, "Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database," *Brain*, vol. 131, no. 9, pp. 2443–2454, 2008.

[47] R. Jena, P. Chaudhari, and J. C. Gee, "Fireants: Adaptive riemannian optimization for multi-scale diffeomorphic matching," *arXiv preprint arXiv:2404.01249*, 2024.

[48] B. Kim and J. C. Ye, "Diffusion deformable model for 4d temporal medical image generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 539–548.

[49] G. Pombo, R. Gray, M. J. Cardoso, S. Ourselin, G. Rees, J. Ashburner, and P. Nachev, "Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models," *Medical Image Analysis*, vol. 84, p. 102723, 2023.

[50] L. Puglisi, D. C. Alexander, and D. Ravì, "Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 173–183.

[51] T. Xia, A. Chartsias, C. Wang, S. A. Tsaftaris, A. D. N. Initiative *et al.*, "Learning to synthesise the ageing brain without longitudinal data," *Medical Image Analysis*, vol. 73, p. 102169, 2021.

[52] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational bayes," 2013.

[53] M. Lorenzi, N. Ayache, G. B. Frisoni, and X. Pennec, "Lcc-demons: A robust and accurate symmetric diffeomorphic registration algorithm," *NeuroImag e*, vol. 81, pp. 470–483, 2013, accessed 20 April 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1053811913004825

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[56] T. N. Wolf, S. Pölsterl, C. Wachinger, A. D. N. Initiative *et al.*, "Daft: A universal module to interweave tabular data and 3d images in cnns," *NeuroImage*, vol. 260, p. 119505, 2022.

[57] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[58] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Springer, 2018, vol. 11070, pp. 729–738, accessed 20 April 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00928-1_82

[59] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[60] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[61] K. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[62] E. L. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[63] A. L. Young and et al., "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference," *Nature Communications*, vol. 9, no. 1, p. 4273, 2018.

[64] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," *International Conference on Machine Learning*, pp. 4114–4124, 2019.

[65] H. M. Fonteijn and et al., "An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease," *NeuroImage*, vol. 60, no. 3, pp. 1880–1889, 2012.

[66] S. Han, S. Yu, M. Shi, M. Harada, J. Ge, J. Lin, C. Prehn, A. Petrera, Y. Li, F. Sam *et al.*, "Leopard: missing view completion for multi-timepoint omics data via representation disentanglement and temporal knowledge transfer," *Nature Communications*, vol. 16, no. 1, p. 3278, 2025.

[67] M. Bilgel and et al., "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging," *NeuroImage*, vol. 134, pp. 658–670, 2016.

[68] A. Chartsias *et al.*, "Factorised graph representation learning: A case study on disease progression modeling," *Medical Image Analysis*, vol. 58, p. 101544, 2019.

[69] E. Dupont, "Learning disentangled joint continuous and discrete representations," *arXiv preprint arXiv:1804.07261*, 2018.

[70] H. Bhatia *et al.*, "Biological image synthesis for functional inference: from structure to function," *Nature Methods*, vol. 18, no. 3, pp. 292–295, 2021.

[71] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101684, 2020.

[72] A. L. Yuille and D. Kersten, "Vision as bayesian inference: analysis by synthesis?" *Trends in cognitive sciences*, vol. 10, no. 7, pp. 301–308, 2006.

[73] R. Gao, M. Wang, and D. Tao, "Deep generative models for missing data imputation in clinical data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9404–9411.

[74] C. R. Jack Jr, R. C. Petersen, Y. C. Xu, P. C. O'Brien, S. C. Waring, E. G. Tangalos, G. E. Smith, R. J. Ivnik, S. N. Thibodeau, and E. Kokmen, "Hippocampal atrophy and apolipoprotein e genotype are independently associated with alzheimer's disease," *Annals of neurology*, vol. 43, no. 3, pp. 303–310, 1998.

[75] B. H. Ridha *et al.*, "Tracking atrophy progression in familial alzheimer's disease: a serial mri study," *The Lancet Neurology*, vol. 7, no. 10, pp. 829–836, 2008.

[76] A. Saykin, H. Wishart, L. Rabin, R. Santulli, L. Flashman, J. West, T. McHugh, and A. Mamourian, "Older adults with cognitive complaints show brain atrophy similar to that of amnestic mci," *Neurology*, vol. 67, no. 5, pp. 834–842, 2006.

[77] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta neuropathologica*, vol. 82, no. 4, pp. 239–259, 1991.

[78] G. B. Karas *et al.*, "Global and local gray matter loss in mild cognitive impairment and alzheimer's disease," *Neuroimage*, vol. 23, no. 2, pp. 708–716, 2004.

[79] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad: results from adni," *Neuroimage*, vol. 44, no. 4, pp. 1415–1422, 2009.

[80] M. Bilgel *et al.*, "A multidimensional spatiotemporal model of alzheimer's disease progression," *Neuroimage*, vol. 131, pp. 193–207, 2016.

[81] B. M. Jedynak *et al.*, "A computational neurodegenerative disease progression score: Method and results with the alzheimer's disease neuroimaging initiative cohort," *Neuroimage*, vol. 63, no. 3, pp. 1478–1486, 2012.

[82] C. R. Jack, R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, E. G. Tangalos, and E. Kokmen, "Hippocampal atrophy and the diagnosis of alzheimer disease," *Archives of neurology*, vol. 54, no. 8, pp. 937–942, 1997.

[83] B. A. Gutman, Y. Wang, A. W. Toga, and P. M. Thompson, "Deformable shape registration with large deformation via lddmm and geodesic shooting," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1215–1229, 2009.

[84] H. Wang, J. H. Suh, S. R. Das, J. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2010.

[85] D. Chan, N. C. Fox, R. I. Scahill, W. R. Crum, J. Barnes, and M. N. Rossor, "Temporal lobe atrophy on mri in normal aging and early alzheimer's disease," *Annals of neurology*, vol. 49, no. 5, pp. 733–742, 2001.

[86] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman *et al.*, "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.

[87] R. Desikan, B. Fischl, H. Cabral, T. Kemper, C. Guttmann, D. Blacker, B. Hyman, M. Albert, and R. Killiany, "Mri measures of temporoparietal regions show differential rates of atrophy during prodromal ad," *Neurology*, vol. 71, no. 11, pp. 819–825, 2008.

[88] Y. Iturria-Medina, R. C. Sotero, P. J. Toussaint, J. M. Mateos-Pérez, and A. C. Evans, "Early role of vascular dysregulation on late-onset alzheimer's disease based on multifactorial data-driven analysis," *Nature Communications*, vol. 7, p. 11934, 2016.