Align-cDAE: Attention-Aligned Conditional Diffusion Auto-encoder for Alzheimer's Disease Progression

Ayantika Das[1] DASAYANTIKA486@GMAIL.COM
[1] *Indian Institute of Technology Madras (IITM)*
Keerthi Ram[2] KEERTHI@HTIC.IITM.AC.IN
[2] *Sudha Gopalakrishnan Brain Centre (SGBC), IITM*
Mohanasankar Sivaprakasam[1,2] MOHAN@EE.IITM.AC.IN

Editors: Under Review for MIDL 2026

Abstract

The integration of multi-modal conditioning with diffusion modeling approaches has been shown to be effective in image-to-image translation tasks. Existing mechanisms usually integrate conditioning information from other modalities by projecting it into a feature space compatible with the image representations. Although these strategies yield improved performance, they do not ensure that information from non-imaging conditioning modalities meaningfully aligns with image features and precisely modulates the generated outputs. In order to better incorporate information from other modalities, we propose a diffusion auto-encoder-based framework for disease progression modeling that explicitly focuses on conditional alignment. This alignment is introduced by constraining the attention between (i) the conditioning attributes and (ii) the feature representations of the model, to focus on the regions exhibiting progression-related changes. This constraint consequently shifts the attention of different layers of the model towards progression-specific regions, generating the required precise anatomical changes. Further, the diffusion auto-encoding-based formulation provides latent representations of images that are compact in nature and suitable for integration of conditions. We have experimentally validated the performance of our model by evaluating on Alzheimer's disease progression generation through various image-level metrics and volumetric assessments. These results demonstrate that enforcing conditional alignment within a diffusion auto-encoding framework leads to more anatomically precise modeling of Alzheimer's disease progression.

Keywords: Multi-modal Conditions, Conditional Alignment, Denoising Diffusion Model, Progression Modeling, Alzheimer's Disease.

1. Introduction

Diffusion-based generative frameworks have emerged as an efficient mechanism for modeling complex image distributions and synthesizing meaningful images (Ho et al., 2020),(Song et al., 2021). These approaches iteratively model the image distribution from pure Gaussian noise, implicitly decomposing image semantics across multiple denoising steps. This multi-step process enables controlled conditional generation, which have been widely leveraged for image synthesis across various tasks (Zhang et al., 2023),(Mou et al., 2024). Specifically, multi-modal conditioning with images and other non-imaging modalities has been effective in target-specific image synthesis like image-to-image translation (Tumanyan et al., 2023) in medical applications (Puglisi et al., 2024). Although these approaches have shown improved performance, most methods incorporate multi-modal information only by ensuring that

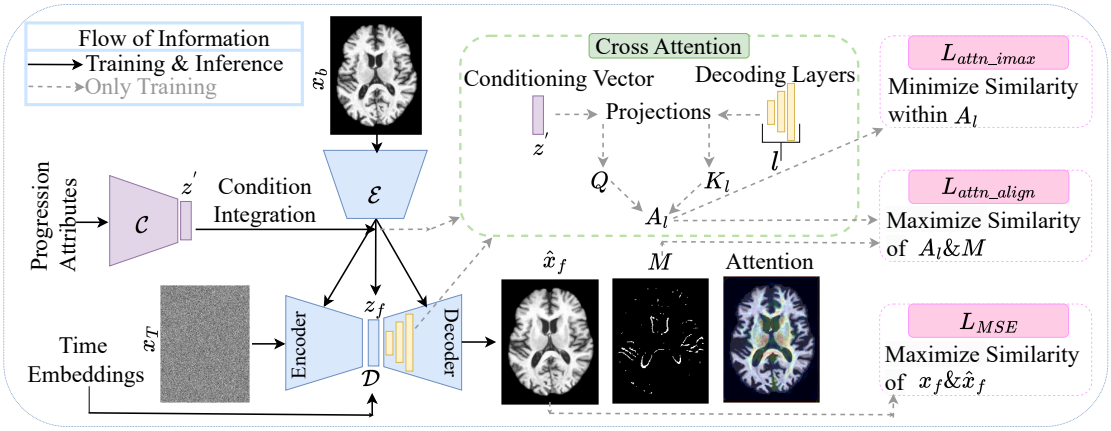

Figure 1: Left to right: The condition encoder ($\mathcal{C}$) integrates progression information with the latent representation of baseline image (x_b) produced by the encoding component ($\mathcal{E}$), guiding the denoising decoder ($\mathcal{D}$) to generate follow-up image ($\hat{x}_f$). Cross-attention (A_l) is extracted between the conditioning vector and decoder layers, and the objective functions enforce alignment of A_l with the progression-specific mask (M).

integration within the feature space is compatible, which does not necessarily introduce a meaningful impact in the image generation. *How do we ensure that multi-modal conditioning precisely modulates the generated images?*

The task of modeling disease progression and predicting longitudinal follow-up images from baseline images can be formulated as an image-to-image translation problem conditioned on various progression attributes. Diffusion-based frameworks have been explored for this task, primarily through latent diffusion models that integrate conditioning directly within the latent denoising process (Puglisi et al., 2024),(Kapoor et al., 2024). Although effective in modeling progression patterns, these methods offer limited control over anatomically precise outputs, since denoising and conditioning occur entirely in the latent space. To improve spatial control during generation, image-diffusion-based techniques have been employed for progression modeling (Yoon et al., 2023; Litrico et al., 2024). However, existing image-space diffusion approaches generally incorporate progression conditioning only by ensuring compatibility within the denoising feature space, without enforcing alignment between the conditioning information and the corresponding image representations. *How do we better align progression conditions with image feature representations of the denoising diffusion model to generate progression-specific anatomical changes?*

To address this, we propose an image-diffusion framework enforcing conditional alignment between progression attributes and image feature maps via an attention-based constraint, which guides the model to focus specifically on regions exhibiting progression-related changes. Furthermore, we employ a diffusion auto-encoding formulation for more effective condition integration during longitudinal prediction, providing a compact latent represen-

tational space (Hudson et al., 2024) suitable for this integration. Our contributions are as follows: **(i)** We introduce Align-cDAE, an image-diffusion framework that enforces ***conditional alignment*** between progression attributes and image feature representations to better capture disease progression-related anatomical changes. **(ii)** We incorporate a ***diffusion auto-encoding*** formulation to integrate baseline image information and progression conditioning within the denoising trajectory, leveraging this mechanism for effective conditional integration. **(iii)** We demonstrate the efficacy of our approach on longitudinal Alzheimer's disease progression MRI datasets, using image-level and volumetric evaluations along with ablation studies, showing that conditional alignment effectively directs the attention of the model toward progression-specific anatomical changes.

Related Works: State-of-the-art conditional generative frameworks model disease progression by integrating progression attributes into the feature representation space and predicting follow-up scans using adversarial (GAN-based) or variational (VAE-based) objectives. GAN-based approaches such as 4D-DANI-Net (Ravi et al., 2022), Identity-cGAN (Jung et al., 2021), IPGAN (Xia et al., 2021), Identity-3D-cGAN (Jung et al., 2023), and SIT-GAN (Wang et al., 2023) incorporate conditioning either directly in the latent space or via feature concatenation, and optimize with biologically informed, identity-preserving, or age-regression based constraints to synthesize follow-up images from baseline scans.

Conditional diffusion-based approaches have emerged as effective tools for modeling longitudinal progression. Latent diffusion models such as BrLP (Puglisi et al., 2024) and MRExtrap (Kapoor et al., 2024) integrate multi-modal conditioning and auxiliary constraints within the latent space, relying on external encoder–decoder mechanisms to transition back to the image domain. Image-space diffusion models, which perform denoising directly on images to model progression, can waive this dependency. Methods such as SADM (Yoon et al., 2023) and TADM (Litrico et al., 2024) incorporate temporal conditioning into the denoising process to predict follow-up scans. Although effective, these approaches primarily introduce multi-modal conditioning without explicitly ensuring that the conditioning information is aligned and precisely modulates the generated images.

2. Methodology

We introduce Align-cDAE, a conditional diffusion auto-encoder-based approach for longitudinal disease progression modeling, focusing on the alignment of progression conditions with the feature representations of the model to better capture progression-related changes. The alignment is incorporated through (i) cross-attention between the conditioning attributes and layer-wise feature representations, and further (ii) constraining the cross-attention to emphasize on structural changes associated with disease progression. The conditional diffusion auto-encoder and the alignment mechanisms are detailed below in the Subsections 2.1 and 2.2. The architectural flow of our method is given in Figure 1.

2.1. Conditional Diffusion Auto-encoder (cDAE)

Our progression-modeling framework is built on a diffusion auto-encoding (DAE) architecture (Preechakul et al., 2022) composed of an encoder ($\mathcal{E}$) and a denoising decoder ($\mathcal{D}$). The model learns the underlying image distribution by progressively transforming pure Gaussian

noise ($x_T \in \mathbb{R}^{H \times W}$) toward the target image ($x_0 \in \mathbb{R}^{H \times W}$) through a time-dependent denoising process, guided by the encoded representation from $\mathcal{E}$. This formulation encourages $\mathcal{E}$ to learn a latent space that captures semantically meaningful image features, while $\mathcal{D}$ reconstructs fine-scale and high-frequency anatomical details (Hudson et al., 2024). The resulting latent representation enables conditioning of the generative process to model disease progression at future time points.

The goal of the ***denoising component*** is to iteratively model the reverse denoising process ($p_\phi(x_{t-1}|x_t)$), estimating x_{t-1} by predicting $\hat{x}_0 = \mathcal{D}(x_t, t)$, given x_t at each diffusion step t. The noisy image x_t is estimated through the forward diffusion process approximated as, $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and α_t is the noise coefficient. The reverse process to be modeled ($p_\phi(x_{t-1}|x_t)$) is approximated,

$$q(x_{t-1}|x_t, \hat{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\left(\frac{x_t - \sqrt{\alpha_t}\hat{x}_0}{\sqrt{1-\alpha_t}}\right), 0\right) \tag{1}$$

The decoding component adopts a U-Net architecture following the standard denoising diffusion formulation. The ***encoding component*** ($\mathcal{E}$) maps baseline images (x_b) into latent representations producing embeddings ($z_b = \mathcal{E}(x_b), z_b \in \mathbb{R}^d$), which guide the denoising decoding process ($\hat{x}_b = \mathcal{D}(x_t, t, \mathcal{E}(x_b))$), formulating an auto-encoding structure. Architecturally, the encoding component ($\mathcal{E}$) is similar to the U-Net encoder in $\mathcal{D}$, excluding the time-dependency.

Conditioning: In order to generate disease progressed follow-up images (x_f) from baseline image (x_b), the latent representations (z) from $\mathcal{E}$ are conditioned with progression attributes (age and disease state). These latent representations are compact and explicit in nature enabling progression conditioning though simple linear transformation, yielding the follow-up prediction as, $\hat{x}_f = \mathcal{D}(x_t, t, z_f)$, where $z_f = z_b + z^{'}$, $z^{'} \in \mathbb{R}^{d^{'}}$ is the conditioning vector. The conditioning vector ($z^{'}$) is obtained by encoding the progression attributes through the condition encoder ($\mathcal{C}$). Architecturally, $\mathcal{C}$ consists of fully connected layers that map the progression attributes (age and disease state, expressed as one-hot vectors) to $z^{'}$.

Training and Inference: During training, the conditional DAE (cDAE) model is optimized over T diffusion steps while enforcing alignment of the model's feature representations with the progression conditions (Align-cDAE), thereby focusing on anatomical regions affected by disease progression. This conditional alignment strategy is detailed in Subsection 2.2. During inference, Align-cDAE predicts follow-up images $\hat{x}_f$ through an iterative denoising process across T_s diffusion steps, guided by the latent representation of the baseline image x_b and conditioned by the progression attributes encoded into the conditioning vector.

2.2. Conditional Alignment

The conditional alignment is achieved by (i) extracting cross-attention between the conditioning vector $z^{'}$ and the layer-wise feature representations ($l \in [1, 2, ..L]$) in the decoder of $\mathcal{D}$ (Vilouras et al., 2025), and (ii) constraining this attention to emphasize regions exhibiting progression-specific changes. The cross-attention mechanism and the associated objective functions for constraining the attention are detailed in Subsections 2.2.1 and 2.2.2.

2.2.1. Cross Attention

Cross-attention is computed between the conditioning vector $z^{'} \in \mathbb{R}^{1\times d^{'}}$ and the first three decoder layers $K_l \in \mathbb{R}^{h\times w\times d_k}$ of $\mathcal{D}$, where d_k, h, and w denote the feature depth and spatial dimensions. The vector $z^{'}$ is projected to match the dimensionality of the decoder features, producing $Q \in \mathbb{R}^{d^{'}\times d_k}$, using fully connected layers. Cross-attention for each layer is then computed as $A_l = (QK_l^T)/\sqrt{d_k}$, yielding $A_l \in \mathbb{R}^{d^{'}\times s}$ with $s = h \times w$. The resulting attention maps are subsequently normalized and passed through a softmax function.

2.2.2. Objective functions

To optimize Align-cDAE, the layer-wise cross-attention maps A_l are aligned to a progression-specific mask that highlights differences between the follow-up x_f and the baseline images x_b. While enforcing A_l to match this progression mask, diversity within the attention maps need to be preserved to maximize information content within feature representations. In addition to these, the original DAE reconstruction objective is retained through an MSE loss, $\mathcal{L}_{MSE} = |x_f - \hat{x}_f|_2^2$. The attention alignment and information maximization objectives are detailed in the following subsections.

Attention Alignment: The attention alignment is enforced by matching the layer-wise cross-attention maps A_l to a progression-specific mask M that highlights changes associated with disease progression. The mask M is constructed by computing the residual between the baseline image x_b and the follow-up image x_f. For similarity maximization, the alignment objective maximizes cosine similarity between A_l and M, which requires both to be spatially compatible. Thus, A_l is averaged across the $d^{'}$ channels and reshaped from s to $h \times w$, yielding $A_l^{'} \in \mathbb{R}^{h\times w}$, while M is down-sampled to $M^{'} \in \mathbb{R}^{h\times w}$. The resulting alignment loss is defined as

$$\mathcal{L}_{attn\_align} = \frac{1}{L}\left(\sum_{l=1}^{L}\left(1 - \cos(A_l^{'}, M^{''})\right)\right) \tag{2}$$

where L denotes the number of decoder layers considered for attention alignment.

Attention Information Maximization: The information content within the cross-attention maps A_l is enhanced by encouraging diversity across their channel dimensions. This is achieved by reducing the squared cosine similarity between different channels of A_l, leading to the following objective formulation,

$$\mathcal{L}_{attn\_imax} = \frac{1}{L}\left(\sum_{l=1}^{L}\sum_{i=1}^{d^{'}}\sum_{\substack{j=1\\ j\neq i}}^{s}\cos^2(A_l^i, A_l^j)\right) \tag{3}$$

Hence, the overall objective function is as follows, $\mathcal{L} = \lambda_1\mathcal{L}_{attn\_imax} + \lambda_2\mathcal{L}_{attn\_align} + \lambda_3\mathcal{L}_{MSE}$, where λ_1, λ_2 and λ_3 are weightage to the losses.

3. Experimental Setup

Dataset and Evaluation Metrics: We conduct experiments using longitudinal T1-weighted brain MRI scans from Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr

Table 1: Quantitative evaluations of Align-cDAE with the baseline methods in terms of image-level metrics (PSNR, SSIM, MSE). Statistical significance ($p < 0.01$) is marked with (*).

Methods	**PSNR (dB) (↑)**	**SSIM (↑)**	**MSE (↓)**
	CN/ MCI & AD	**CN/ MCI & AD**	**CN/ MCI & AD**
Naive Baseline	27.25 ± 2.12/ 26.75 ± 2.07	0.93 ± 0.021/ 0.92 ± 0.021	0.0021 ± 0.001/ 0.0024 ± 0.001
IPGAN (Xia et al., 2021)	25.86 ± 2.12/ 25.31 ± 2.13	0.92 ± 0.032/ 0.91 ± 0.034	0.0030 ± 0.001/ 0.0034 ± 0.002
BrLP (Puglisi et al., 2024)	26.71 ± 1.02/ 26.20 ± 1.14	0.79 ± 0.022/ 0.79 ± 0.025	0.0029 ± 0.001/ 0.0030 ± 0.001
DE-CVAE (He et al., 2024)	27.32 ± 2.98/ 26.99 ± 2.83	0.65 ± 0.090/ 0.63 ± 0.082	0.0023 ± 0.001/ 0.0024 ± 0.001
SITGAN (Wang et al., 2023)	28.73 ± 3.25/ 28.09 ± 3.23	0.94 ± 0.033/ 0.93 ± 0.034	0.0019 ± 0.001/ 0.0022 ± 0.001
cDAE	28.83 ± 4.23/ 28.27 ± 4.81	0.94 ± 0.038/ 0.93 ± 0.050	0.0019 ± 0.001/ 0.0021 ± 0.002
Align-cDAE	**29.82* ± 3.50/ 29.52* ± 3.82**	**0.95* ± 0.031/ 0.94* ± 0.034**	**0.0018* ± 0.001/ 0.0019* ± 0.001**

et al., 2008), including subjects diagnosed as cognitively normal (CN), mildly cognitively impaired (MCI), or with Alzheimer's disease (AD). Images span ages 63–87 and include all genders. For each subject, baseline–follow-up pairs were constructed, with an average inter-scan interval of 2.93 ± 1.35 years.

Image-level Metrics: We assess the similarity between generated follow-up ($\hat{x}_f$) and ground-truth follow-up (x_f) images using: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Squared Error (MSE). *Volume-level Metrics*: To evaluate progression modeling in 3D, we compute region-wise Mean Absolute Error (MAE) between $(V^r_{\hat{X}_f} - V^r_{X_b})/(V^r_{X_b})$ and $(V^r_{X_f} - V^r_{X_b})/(V^r_{X_b})$, where V^r_X denotes the voxel count of region r ($r \in$ Hippocampus/ Amygdala/ Lateral Ventricular) extracted from each 3D volume (X) via segmentation using SynthSeg (Billot et al., 2023).

Implementation Details: *Dataset Details*: From the ADNI cohort, we curated a *Training set* of 486 subjects (179 CN, 160 MCI, 147 AD) and a *Test set* of 466 subjects (159 CN, 156 MCI, 151 AD). The pre-processed steps followed were skull stripping (Isensee et al., 2019), affine registration to MNI space, and intensity normalization (Shinohara et al., 2014). *Model Details*: The models were implemented in PyTorch version 2.0.1 on an 80 GB NVIDIA A100 GPU and CUDA Version: 12.1. The parameter specifications are detailed in Appendix A.1. *Baseline Methods*: We have compared with generative models used for progression modeling, including GAN-based approaches (IPGAN (Xia et al., 2021),

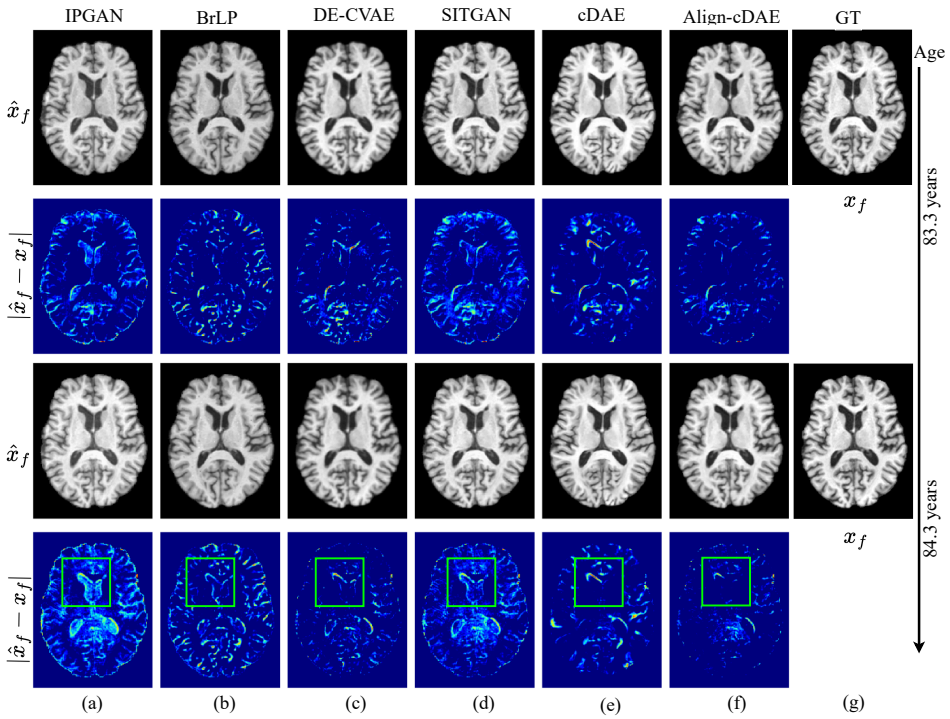

Figure 2: Left to right: columns (a)-(f) present the method-wise comparison of predicted follow-up images ($\hat{x}_f$) and their respective errors with the ground truth (x_f) in column (g). Top to bottom: Upper and lower two rows present results of 83.3 and 84.3 years of age, respectively, from the AD disease category. The green box highlights that Align-cDAE is able to better predict progression-related changes.

SITGAN (Wang et al., 2023)), VAE-based method (DE-CVAE (He et al., 2024)), and diffusion-based approach (BrLP (Puglisi et al., 2024)). Additionally, an ablated version of Align-cDAE trained only with the MSE loss, without conditional alignment was included as cDAE.

4. Results and Discussion

4.1. Quantitative and Qualitative Analysis

The quantitative evaluations comparing Align-cDAE with baseline models are reported in Table 1, using image-level metrics (PSNR, SSIM, and MSE) to compare predicted ($\hat{x}_f$) with ground truth (x_f) follow-up image. As shown, Align-cDAE achieves relatively better performance due to the explicit incorporation of the attention alignment mechanism. As compared to the DAE-based methods, SITGAN performs relatively lower, since the DAE approaches explicitly learn progression-related anatomical changes, while SITGAN relies on an age estimation mechanism to implicitly learn these changes. Among the remaining baselines, the

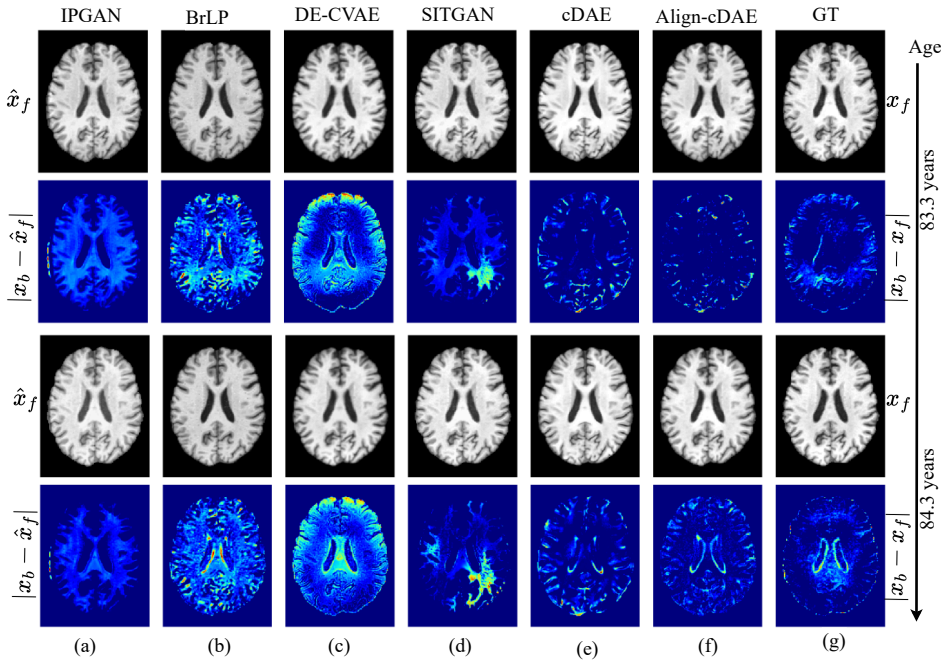

Figure 3: Left to right: columns (a)-(f) present the method-wise comparison of predicted follow-up images ($\hat{x}_f$) and their respective difference maps ($|x_b - \hat{x}_f|$) with the ground truth baseline image (x_b). The last column (g) highlights ground truth follow-up and differences ($|x_b - x_f|$). Top to bottom: Upper and lower two rows present results of 83.3 and 84.3 years of age, respectively, from the AD disease category. The difference maps highlight that the progression-related anatomical changes of Align-cDAE more resemble the ground-truth.

VAE-based (DE-CVAE) and VAE–latent-diffusion (BrLP) approaches show relatively lower performance, since these models tend to produce blurrier structural details, leading to more image-level errors despite incorporation of better constraints for modeling disease progression. The GAN-based approach (IPGAN) also employs an implicit mechanism to transform baseline to follow-up images, but lacks awareness about progression specific changes in the image. Overall, Align-cDAE ***performs better*** across baselines, demonstrating the effectiveness of conditional alignment that helps to ***focus*** on ***progression-specific*** regions for controlled image generations.

Figure 2 presents a qualitative comparison of absolute error between predicted and ground-truth follow-up images ($|\hat{x}_f - x_f|$) for two future time points (83.3 and 84.3 years) of a subject aged 80.8 years at baseline (AD group). Align-cDAE shows relatively lower error than all baselines. Compared with the baseline, cDAE, Align-cDAE produces smaller errors in regions associated with disease progression, such as the lateral ventricles. SITGAN exhibits higher error than the DAE-based methods, particularly at 84.3 years, indicating

Table 2: Volumetric assessment of Align-cDAE with baseline methods in terms of MAE in different anatomical regions. Statistical significance ($p < 0.01$) is marked with (*).

Methods	**MAE (↓)**		
	Hippocampus	**Amygdala**	**Lateral Ventricles**
IPGAN (Xia et al., 2021)	0.3348 ± 0.0338	0.3283 ± 0.0260	0.5445 ± 0.5240
BrLP (Puglisi et al., 2024)	0.1960 ± 0.0552	0.1731 ± 0.0529	0.3702 ± 0.1012
DE-CVAE (He et al., 2024)	0.1871 ± 0.0727	0.1183 ± 0.1117	0.1747 ± 0.1470
SITGAN (Wang et al., 2023)	0.1161 ± 0.0288	0.0211 ± 0.0201	0.1436 ± 0.1053
cDAE	0.0550 ± 0.0500	0.0200 ± 0.0400	0.1250 ± 0.1200
Align-cDAE	**0.0282*** ± **0.0203**	**0.0199*** ± **0.0168**	**0.0705*** ± **0.0261**

that progression related factors are not effectively captured. The VAE-based methods (DE-CVAE and VAE–latent-diffusion) capture progression-relevant information but generate blurrier outputs, increasing error in high-frequency, and subject-specific structures. IPGAN produces higher error, affecting both progression-sensitive and general anatomical areas. Overall, Align-cDAE better captures progression-specific anatomical changes, ***reducing error*** in both progression-related and ***identity-preserving*** regions.

Analysis of Progression-related Changes with Age: Figure 3 compares anatomical changes produced by different methods relative to baseline images (age 80.8 years) for two follow-up time points (83.3 and 84.3 years). The ground-truth difference maps show an age-based progression hierarchy, with more pronounced ventricular expansion at 84.3 years due to accelerated atrophy around the ventricles. Align-cDAE better reproduces this ventricular growth pattern and preserves the expected progression hierarchy. The DAE baseline (cDAE) captures ventricular enlargement but lacks the precision needed to localize progression-specific changes. GAN-based approaches (SITGAN and IPGAN) produce outputs that remain close to the baseline image, not precisely highlighting the expected progression in the ventricular region. VAE and latent-diffusion VAE methods introduce broader, less localized differences both within and outside progression-sensitive regions. Overall, the difference map of Align-cDAE indicates that the design specification introduced to focus on progression-related changes ***facilitates*** generation of the ***required anatomical*** changes.

4.2. Volumetric Analysis

To evaluate how well progression-related changes are captured in 3D, we compare normalized volumetric changes between predicted ($V_{\hat{X}_f}$) and ground-truth follow-up (V_{X_f}) volumes relative to the baseline (V_{X_b}). The MAE between $(V^r_{\hat{X}_f} - V^r_{X_b})/(V^r_{X_b})$ and $(V^r_{X_f} - V^r_{X_b})/(V^r_{X_b})$ for predicted and ground truth follow-ups respectively are reported in Table 2, for three anatomical regions (r). Align-cDAE achieves relatively lower errors across all regions, consistent with the image-level results where DAE-based methods are better than other baselines. GAN-based approaches (SITGAN and IPGAN) rely on implicit age constraints and do not to capture region-specific anatomical changes precisely, resulting in higher errors.

VAE and latent-diffusion VAE models produce non-localized volumetric differences due to blurrier high-frequency reconstructions, limiting their ability to model precise progression. Overall, Align-cDAE shows ***lower volumetric error*** by concentrating anatomical changes in progression-relevant regions, producing ***meaningful 3D*** progressions suitable for downstream analyses.

4.3. Analysis of Attention Alignment

In order to assess the impact of attention alignment into our modeling approach, we extract the mid-layer ($l = 10$, not considered for alignment loss) of the decoder of $\mathcal{D}$ and visualize the attention map by averaging across its channel dimension. Figure 4 compares the attention maps of Align-cDAE and cDAE from layers not considered for loss computation, alongside their corresponding difference maps highlighting generated changes. (i) *Attention Maps*: From the figure, it is evident that Align-cDAE focuses attention around progression-relevant regions, particularly the ventricles, whereas cDAE shows a diffuse attention pattern. (ii) *Difference Maps*: The difference maps (between the ground-truth baseline and predicted follow-up $(\hat{x}_f - x_b)$) produced by Align-cDAE more closely resemble the actual structural changes seen in the ground-truth $(x_f - x_b)$. This suggests that Align-cDAE ***better captures progression*** patterns.

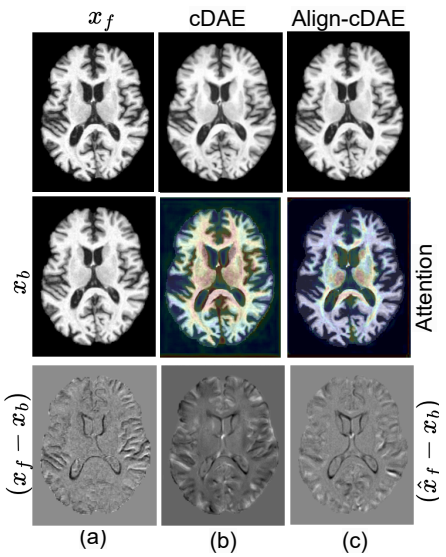

Figure 4: Left to right: (a) Ground truth, (b) cDAE, and (c) Align-cDAE. Top to bottom: (i) Ground-truth and predicted follow-up images, (ii) (a) Ground truth baseline and (b)-(c) model attention maps overlaid on images, and (iii) Difference between ground-truth/ predicted follow-up and ground-truth baseline images.

5. Conclusion

We have introduced a diffusion auto-encoding framework that enforces alignment between multi-modal conditioning and image features, enabling precise modulation of longitudinal disease progression synthesis. The auto-encoding formulation provides a compact latent space enabling effective condition integration. We evaluated the model against multiple baselines using image-level and volumetric metrics on an Alzheimer's disease progression dataset. Ablation studies further demonstrate that enforcing conditional alignment steers the decoder layers to focus attention on progression-specific anatomical regions. *Limitations and Future Scope:* The current model incorporates limited progression attributes (age and disease state) for conditioning, which can be extended to richer clinical modalities that could further enhance the modeling of progression patterns. Overall, our findings highlight that it is essential for conditional models to focus on ***alignment of multi-modal information*** for achieving precise image generation.

Acknowledgments

We would like to thank the entire team at Sudha Gopalakrishnan Brain Centre, IITM, for their consistent support. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

References

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

Rosemary He, Gabriella Ang, Daniel Tward, and Alzheimer's Disease Neuroimaging Initiative. Individualized multi-horizon mri trajectory prediction for alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 26–37. Springer, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick 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*, 27(4):685–691, 2008.

Euijin Jung, Miguel Luna, and Sang Hyun Park. Conditional gan with an attention-based generator and a 3d discriminator for 3d medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–328. Springer, 2021.

Euijin Jung, Miguel Luna, and Sang Hyun Park. Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. *Pattern Recognition*, 133:109061, 2023.

Jaivardhan Kapoor, Jakob H Macke, and Christian F Baumgartner. Mrextrap: Linear prediction of brain aging in autoencoder latent space of mri scans. In *Medical Imaging with Deep Learning*, 2024.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. Tadm: Temporally-aware diffusion model for neurodegenerative progression on brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 444–453. Springer, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–183. Springer, 2024.

Daniele Ravi, Stefano B Blumberg, Silvia Ingala, Frederik Barkhof, Daniel C Alexander, Neil P Oxtoby, Alzheimer's Disease Neuroimaging Initiative, et al. Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia. *Medical Image Analysis*, 75:102257, 2022.

Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

Konstantinos Vilouras, Ilias Stogiannidis, Junyu Yan, Alison Q O'Neil, and Sotirios A Tsaftaris. Anatomy-grounded weakly supervised prompt tuning for chest x-ray latent diffusion models. *arXiv preprint arXiv:2506.10633*, 2025.

Clinton J Wang, Natalia S Rost, and Polina Golland. Spatial-intensity transforms for medical image-to-image translation. *IEEE transactions on medical imaging*, 42(11):3362–3373, 2023.

Tian Xia, Agisilaos Chartsias, Chengjia Wang, Sotirios A Tsaftaris, Alzheimer's Disease Neuroimaging Initiative, et al. Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, 73:102169, 2021.

Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pages 388–400. Springer, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3836–3847, 2023.

Appendix A.

A.1. Parameter Details

Table 3: Hyperparameters and configuration details used in our experiments.

Parameters	**Values**	**Parameters**	**Values**
Axial Height (H), Width (W)	208, 160	Inference Diffusion Steps (T_s)	50
Latent Dimension (d)	512	Decoder Layers for Alignment (L)	3
Conditioning Vector Dimension (d')	50	Decoder Layer Dimensions (d_k, h, w)	512, 16, 16
Training Diffusion Steps (T)	1000	Loss Weights ($\lambda_1, \lambda_2, \lambda_3$)	0.01, 0.01, 1
Epochs	180	Optimizer	Adam
Learning Rate	0.001	Diffusion Noise Schedule	Linear