LoCI-DiffCom: Longitudinal Consistency-Informed Diffusion Model for 3D Infant Brain Image Completion

Zihao Zhu[1], Tianli Tao[1], Yitian Tao[1], Haowen Deng[1], Xinyi Cai[1],
Gaofeng Wu[1], Kaidong Wang[1], Haifeng Tang[1], Lixuan Zhu[1], Zhuoyang Gu[1],
Dinggang Shen[1,2,3], and Han Zhang[1]

[1] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[2] Shanghai United Imaging Intelligence Co., Ltd, Shanghai, China
[3] Shanghai Clinical Research and Trail Center, Shanghai, China
{zhuzh2023,zhanghan2}@shanghaitech.edu.cn

Abstract. The infant brain undergoes rapid development in the first few years after birth. Compared to cross-sectional studies, longitudinal studies can depict the trajectories of infants' brain development with higher accuracy, statistical power and flexibility. However, the collection of infant longitudinal magnetic resonance (MR) data suffers a notorious dropout problem, resulting in incomplete datasets with missing time points. This limitation significantly impedes subsequent neuroscience and clinical modeling. Yet, existing deep generative models are facing difficulties in missing brain image completion, due to sparse data and the nonlinear, dramatic contrast/geometric variations in the developing brain. We propose LoCI-DiffCom, a novel Longitudinal Consistency-Informed Diffusion model for infant brain image Completion, which integrates the images from preceding and subsequent time points to guide a diffusion model for generating high-fidelity missing data. Our designed LoCI module can work on highly sparse sequences, relying solely on data from two temporal points. Despite wide separation and diversity between age time points, our approach can extract individualized developmental features while ensuring context-aware consistency. Our experiments on a large infant brain MR dataset demonstrate its effectiveness with consistent performance on missing infant brain MR completion even in big gap scenarios, aiding in better delineation of early developmental trajectories.

Keywords: Medical image generation · Infant brain development · Diffusion model · Magnetic resonance imaging (MRI).

1 Introduction

The brains of human infants undergo dramatic morphometric and geometric changes during early infancy. The total cerebral volume increases from about 30% to 80% of the adult size during the first two years after birth[3, 12]. Besides global changes, local brain areas evolve more significantly[12, 7], laying foundations for emerging cognitive and learning abilities[18]. Recent advancements more and more rely on the use of longitudinal magnetic resonance imaging (MRI) to characterize growth trajectories[8, 23]. Compared to cross-sectional data, longitudinal MRI can unravel developmental trajectories, especially individual differences in these curves, with elevated accuracy, statistical power, and analytic flexibility[13]. However, longitudinal infant MRI faces enormous challenges due to poor cooperation during scanning, heavy imaging noise and artifacts[12], even subject drop-out during follow-up stages, resulting in missing data[26]. Commonly, the age interval of two data from the same infant is too large to reveal dynamic and nonlinear developmental changes in-between. Moreover, the contrast of longitudinal infant brain MRI changes dramatically due to immature myelination[5], further complicating data completion. The field calls for high-fidelity image completion with accuracy for small brain structures that undergo larger changes.

Deep generative models, such as generative adversarial networks (GANs) and diffusion probabilistic models (DPMs)[6, 24], have achieved significant success in the field of image generation. DPMs have particularly shown superiority in generating 3D medical images with rich details[2, 20, 19]. The pioneering studies have attempted to use DPMs for longitudinal image completion[10, 4, 25], but this task remains a cutting-edge challenge for infant brain. For instance, generating deformation fields[10] relies on largely uniform deformation assumption, which does not always hold true for infant brain MRI. Another longitudinal MRI completion study relied on single guidance image[4], which might cause large distortions in the infant scenario. Multiple guiding images will offer a better-controlled condition for DPMs by explicitly using multiple past time points to predict future data[25]. However, the sequence-aware transformer used was borrowed from a video-based vision transformer[1], only suitable for extracting simple temporal features from natural video frames but could fail in more complicated infant MRI completion. Learning longitudinal sequence in early brain development is not as simple as that used in video frame completion, an efficient yet powerful algorithm that can integrate spatiotemporal semantic information from very sparse sequences is urgently needed.

This paper presents Longitudinal Consistency-Informed Diffusion model for infant brain image Completion (LoCI-DiffCom). This novel algorithm can provide adaptive guidance to constrain a conditional DPM for generating high-fidelity missing infant brain MRI data with any paired preceding and subsequent time points of any interval. As the missing data bares dramatic temporal variations and large individual variability, we introduce a longitudinal consistency-informed module that fuses two time-point data to achieve context-aware consistency for carefully guiding DPM-based generation. As enormous spatiotemporal information is involved, to adaptively adjust the significance of various semantic features across spatial and channel domains, we incorporate light-weighted channel-spatial attention[15] into DPM. Our method on completing the Baby Connectome Project (BCP) dataset[8] demonstrates its efficacy in a sparse sequence scenario.

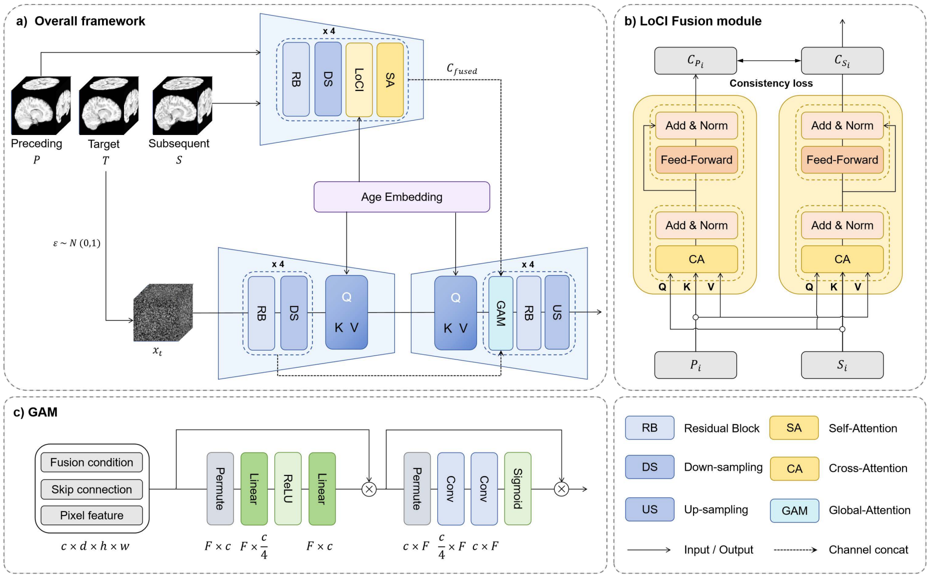

Fig. 1. The overall architecture of our proposed LoCI-DiffCom (a) with Longitudinal Consistency-Informed (LoCI) Fusion module (b) and a global attention mechanism (GAM) (c).

2 Method

Fig. 1a illustrates the architecture of our LoCI-DiffCom. Specifically, a LoCI module, detailed in Fig. 1b, is designed to fuse the preceding and subsequent data to form conditions C_{fused} with context consistency for guiding diffusion model generation. Then, C_{fused} is informed into the DPM via a global attention mechanism (GAM, Fig. 1c) for modulating the importance of various semantic information in the denoising network, enabling the better decoding.

2.1 Longitudinal Consistency-Informed Fusion Module

In conditional DPMs, the generated image largely depends on the congruence and association between the conditions and the target. For infant early brain development, the condition image may differ significantly from the target image due to infants' rapid developmental rate. Therefore, to accurately guide longitudinal image generation, we design a LoCI fusion module to better integrate images collected preceding (P) and subsequent (S) to the missing time point, resulting in semantic features C_{fused} with high relevance to the target. This strategy also maximally utilizes existing data for high-fidelity image completion.

In Fig. 1a. P and S are initially encoded to extract their semantic representations. The encoder consists of n residual blocks and n two-fold downsampling to get n pairs of encoded features of varied sizes, denoted as P_i and S_i. In order to extract the individual's developmental characteristics by integrating the semantic features of preceding and subsequent time points, we input each set of P_i and S_i into the LoCI module (detailed in Fig. 1b). LoCI has a transformer-based architecture with multi-head cross-attention, consisting of three Transformer encoders. The initially encoded features sequentially pass through the three cross-attention transformers with their respective Q, K, V derived from fully connected linear transformations. By switching Q to perform cross-attention, followed by layer normalization and feed-forward network, the features from the preceding time point gradually exchange information with those from subsequent time point, with their commonality enhanced. After LoCI fusion, the initial features turn into fused features with context-aware consistency C_{Pi} and C_{Si}. To achieve more accurate feature fusion, we minimize the Mean Squared Error (MSE) between C_{Pi} and C_{Si}, aiming to reach common characteristics that represents individual's unique developmental traits with high relevance to the target. The loss function, L_{LoCI}, can be formulated as:

$$L_{LoCI} = MSE(C_P, C_S) = \frac{1}{n}\sum_{i=1}^{n}(C_{Pi} - C_{Si})^2, \tag{1}$$

where n is the number of LoCI fusion modules. The output of the LoCI module is C_{Si}, which has already integrated information from preceding time point and is then fed into a final Transformer encoder with self-attention, resulting in fused conditions C_{fused}, as the guidance for the denoising network.

2.2 Hybrid Attention Mechanism in the DPMs

In the implementation of conditional DPMs, it is straightforward that the individual guidance image and the age information are encoded and then simply concatenated, or directly added to the latent space, prior to the denoising process. However, this approach may obfuscate features with different semantics. Instead of doing so, we adopt a hybrid attention mechanism in the implementation of LoCI-DiffCom. That is, we embed the fused image feature C_{fused} and the age information x_{mo} using global attention and cross-attention, respectively. As illustrated in Fig. 1c, the global attention mechanism (GAM) comprises a lightweight channel attention mechanism and spatial attention. In channel attention, a 4D permutation merges spatial dimensions and swaps their order with the channel dimension, followed by a two-layer multi-layer perceptron (MLP) with a channel reduction ratio r to extract a channel attention map. In spatial attention, the original dimension permutation is first restored, followed by two convolutional layers focusing on extracting an attention map for the spatial dimension. This attention is capable of learning significant global information by traversing dimensions across spatial and channel domains. For the age information of the target image x_{mo}, we learn the corresponding token, integrate it with the time step, and use it as a query to compute cross-attention with pixel features at both the encoder and decoder in the denoising network. As for the age tokens of the conditional images, they are simply added to the inputs of the final LoCI module.

2.3 Model Training

The objective function of the conditional DPM is defined as follows:

$$L_{diff} = E[||\epsilon - \epsilon_\theta(x_t; t, x_{mo}, C_{fused})||^2], \tag{2}$$

where ϵ represents a random Gaussian distribution, x_t is the noised target image, t represents the number of time steps for adding noise. Given that C_{fused} is the input to the DPM, the objective function of the DPM L_{diff} will optimize the parameters of the LoCI fusion module θ_{LoCI}. However, L_{LoCI} does not affect the parameters of diffusion θ_{diff}. Our network employs an end-to-end training approach, allowing simultaneous optimization of both parts, formulated as:

$$L = L_{diff} + \lambda L_{LoCI} \tag{3}$$

Given the convergence rate of the LoCI module being greater than that of the denoising network, λ can be selected to modulate the extent of optimization for the LoCI module.

3 Experiments and Results

3.1 Dataset and Implemention

BCP Dataset In our study, we utilized a longitudinal infant MRI dataset from Baby Connectome Project (BCP)[8], comprising a total of 170 infants aged 0-26 months, with 655 T1-weighted structural MR scans. After rigorous quality control, we randomly selected 584 scans from 154 infants for training and 71 scans from 16 different infants for testing. All data were preprocessed according to a standard procedure[14]. Considering the computational cost, input images were resampled to 2×2×2 mm^3.

Implementation Details Our model is implemented using PyTorch[17] trained on an Nvidia A100 GPU with memory of 80 GB. For diffusion, the noise level is set from 10^{-4} to 5×10^{-3} linearly with 1000 steps. Adam optimization[11] is used with a learning rate of 2×10^{-4}. During inference, denoising is performed with 80 skip steps across 1000 steps. The model in our experiment has $n = 4$ LocI modules. $\lambda = 0.6$ is selected to modulate the optimization for LoCI. Channel reduction ratio $r = 4$ in GAM.

3.2 Results

Evaluation Metrics and Baseline Methods We choose several state-of-the-art methods as baseline models for comparison: 1) GAN-based methods: Conditional GAN[16] and Pix2Pix GAN[9], which use a UNet-based GAN model to synthesize the missing image given a reference image as the condition, as well as 2) Diffusion-based methods: Conditional DDIM[4] and SADM[25], which can only utilize images from the previous time points to predict later time point.

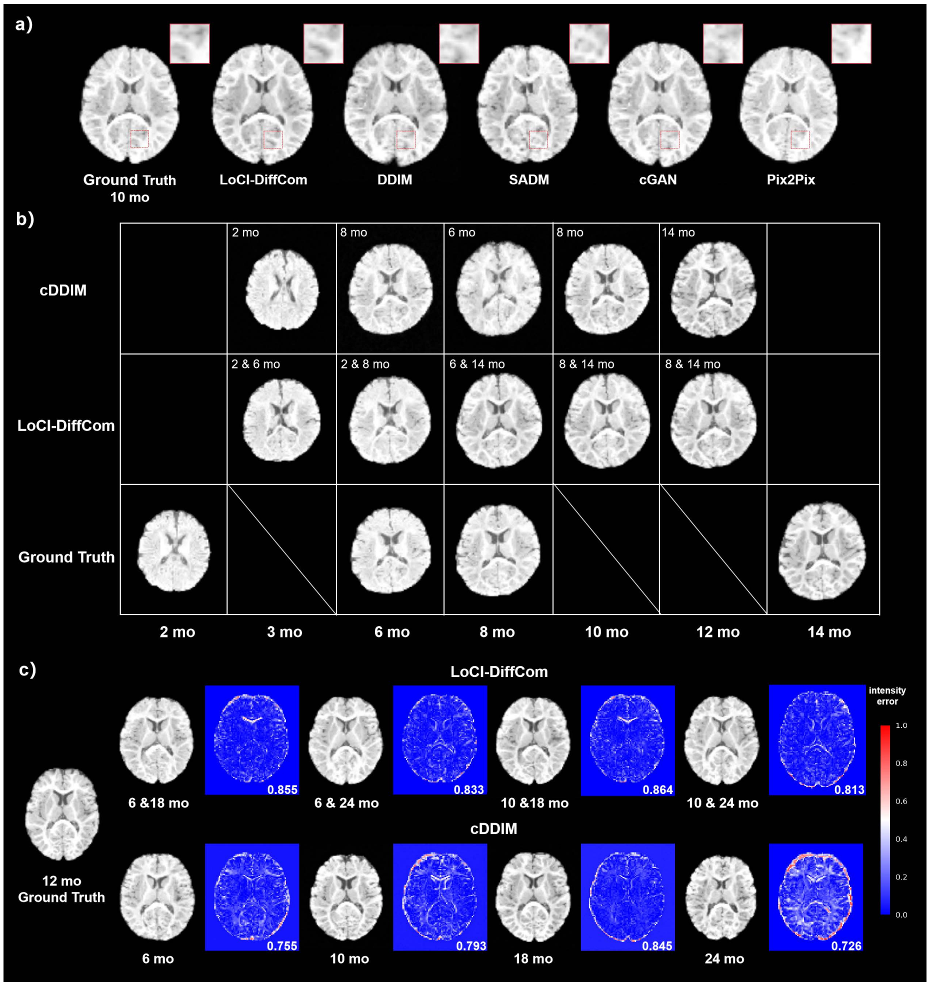

Fig. 2. Qualitative comparison between baseline methods and our proposed LoCI-DiffCom (a). Comparison of longitudinally generative performance, with the text on the top-left representing conditional images utilized (b). Visual comparison and error maps of different longitudinal consistency guidance. The bottom-right of the error map represents SSIM score (c).

Table 1. Quantitative comparison between baselines and our proposed LoCI-DiffCom.

Method	PSNR	SSIM	Dice	
			WM	GM
cGAN	23.73	0.794	0.642	0.629
Pix2Pix	24.01	0.796	0.643	0.622
DDIM	24.07	0.798	0.646	0.627
SADM	23.57	0.781	0.569	0.587
LoCI-DiffCom	**25.52**	**0.845**	**0.656**	**0.650**

The methods are compared using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Furthermore, we evaluate the fidelity of the generated images. Specifically, we employ infant-dedicated brain MRI segmentation[22][27][21] to parcellate brain gray matter and white matter using our generated images and use Dice coefficient to evaluate the segmentation performance. A higher Dice score indicates that the generated image is more reasonable.

Qualitative and Quantitative Comparisons In 2a, the subject has images at 4, 7, 10, and 14 months. We selected the image from the 10 month mark as the ground truth. The GAN-based method and conditional DDIM used a single image at 7 months of age as guidance. SADM was guided by several preceding images at 4 and 7 months of age. Our LoCI-DiffCom utilized the images from both 7 and 14 months as guidance. As shown in Fig. 2a, our method outperformed other methods in terms of image details and preserved more individual-specific information. Quantitative evaluations are summarized in Table. 1. LoCI-DiffCom outperformed GAN-based methods by 5%-10% and was slightly better than the diffusion-based methods. It is worth noting that SADM's results, with preceding time points' images as guidance, were even worse than those using only one guidance image, largely due to the prominent structural diversity in early development.

Since conditional DDIM performed best among the baseline methods, we visually compare its generated longitudinal sequence with ours. The age(s) of guidance image(s) is shown in the top left corner of each sub-figure in Fig. 2b. Our model has better longitudinal consistency, showing a gradual grow-up trend. However, the brain size of the images generated by the conditional DDIM varied in an unreasonable manner.

Ablation Study We examined the impact of the number of our proposed LoCI modules on the model's effectiveness and quantitatively compared the image quality and fidelity. As shown in Table. 2, LoCI (×4) achieves a 3% improvement in SSIM compared to LoCI (×1). We also investigated the effectiveness of our hybrid attention mechanism in the diffusion module. In both cases, our attention mechanism showed a performance improvement.

Table 2. Ablation studies of LoCI-DiffCom.

Method		PSNR	SSIM	Dice	
LoCI	Attention			WM	GM
×4	✓	**25.52**	**0.845**	0.656	**0.650**
×1	✓	24.60	0.806	0.605	0.629
×4		25.05	0.830	**0.661**	0.647
×1		24.20	0.801	0.600	0.615

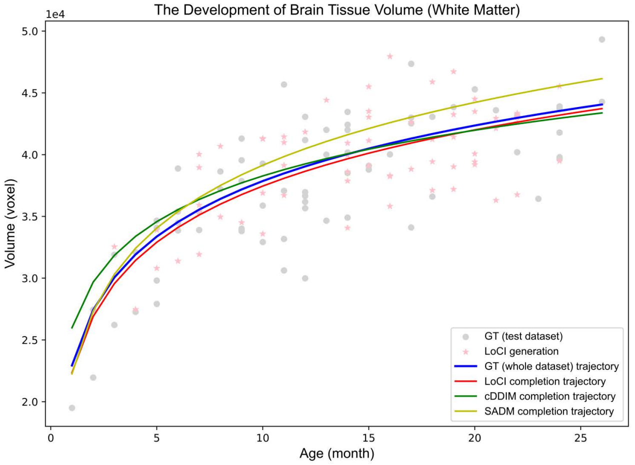

Fig. 3. The longitudinal growth trajectories of infant brain white matter volume.

Analysis of Longitudinal Consistency Guidance The quality of the generated image could highly depend on guidance selection. We investigated the influence of different guidance selection strategies on the generated image by applying images at different ages as guidance to both conditional DDIM and our LoCI-DiffCom. As for generating image at 12 months of age, LoCI-DiffCom with any pair of guidance images imputed data with higher similarities to the ground truth in a highly robust manner (Fig. 2c). Conditional DDIM showed large instability, particularly with guidance far from the target age, where the generation quality was generally poor.

Performance on a Downstream Task: Delineating Developmental Trajectory after Data Completion. We further assessed the fidelity of data completion from a developmental neuroscience perspective by fitting the developmental trajectory of total white matter volume using a linear mixed-effect model with a log-linear function. Fig. 3 demonstrates that the developmental trajectory completed by our method is closer to that of the ground truth, while SADM and conditional DDIM exhibit significant discrepancies with the ground truth trajectory.

4 Conclusion

We designed a novel conditional diffusion model, LoCI-DiffCom, aimed at completing missing infant brain images for better longitudinal studies. The proposed consistency-informed module effectively merges conditions from preceding and subsequent age time points and generates high-fidelity data that preserves longitudinal changes and individual variability. It is also highly stable in extreme scenarios where the guiding images are far from the target. Our approach offers a potential solution to complete missing data for more accurate developmental neuroscience studies.

Acknowledgments. This work is supported by the STI 2030-Major Projects (No.2022 ZD0209000), National Natural Science Found-ation of China (No. 62203355), Shanghai Pilot Program for Basic Research - Chinese Academy of Science, Shanghai Branch (No. JCYJ-SHFY-2022-014), Shenzhen Science and Technology Program (No. KCXFZ 20211020163408012), and Shanghai Pujiang Program (No. 21PJ1421400).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
2. Dorjsembe, Z., Odonchimed, S., Xiao, F.: Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In: Medical Imaging with Deep Learning (2022)
3. Gilmore, J.H., Knickmeyer, R.C., Gao, W.: Imaging structural and functional brain development in early childhood. Nature Reviews Neuroscience **19**(3), 123–137 (2018)
4. Guo, L., Tao, T., Cai, X., Zhu, Z., Huang, J., Zhu, L., Gu, Z., Tang, H., Zhou, R., Han, S., et al.: Cas-diffcom: Cascaded diffusion model for infant longitudinal super-resolution 3d medical image completion. arXiv preprint arXiv:2402.13776 (2024)
5. Hazlett, H.C., Gu, H., McKinstry, R.C., Shaw, D.W., Botteron, K.N., Dager, S.R., Styner, M., Vachet, C., Gerig, G., Paterson, S.J., et al.: Brain volume findings in 6-month-old infants at high familial risk for autism. American Journal of Psychiatry **169**(6), 601–608 (2012)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Holland, D., Chang, L., Ernst, T.M., Curran, M., Buchthal, S.D., Alicata, D., Skranes, J., Johansen, H., Hernandez, A., Yamakawa, R., et al.: Structural growth trajectories and rates of change in the first 3 months of infant brain development. JAMA neurology **71**(10), 1266–1274 (2014)
8. Howell, B.R., Styner, M.A., Gao, W., Yap, P.T., Wang, L., Baluyot, K., Yacoub, E., Chen, G., Potts, T., Salzwedel, A., et al.: The unc/umn baby connectome project (bcp): An overview of the study design and protocol development. NeuroImage **185**, 891–905 (2019)

9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
10. Kim, B., Ye, J.C.: Diffusion deformable model for 4d temporal medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 539–548. Springer (2022)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Knickmeyer, R.C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J.K., Hamer, R.M., Lin, W., Gerig, G., Gilmore, J.H.: A structural mri study of human brain development from birth to 2 years. Journal of neuroscience **28**(47), 12176–12182 (2008)
13. Kraemer, H.C., Yesavage, J.A., Taylor, J.L., Kupfer, D.: How can we learn about developmental processes from cross-sectional studies, or can we? American Journal of Psychiatry **157**(2), 163–171 (2000)
14. Li, G., Nie, J., Wang, L., Shi, F., Lin, W., Gilmore, J.H., Shen, D.: Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age. Cerebral cortex **23**(11), 2724–2733 (2013)
15. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021)
16. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Paterson, S.J., Heim, S., Friedman, J.T., Choudhury, N., Benasich, A.A.: Development of structure and function in the infant brain: Implications for cognition, language and social behaviour. Neuroscience & Biobehavioral Reviews **30**(8), 1087–1105 (2006)
19. Peng, W., Adeli, E., Zhao, Q., Pohl, K.M.: Generating realistic 3d brain mris using a conditional diffusion probabilistic model. arXiv preprint arXiv:2212.08034 (2022)
20. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)
21. Shi, F., Fan, Y., Tang, S., Gilmore, J.H., Lin, W., Shen, D.: Neonatal brain image segmentation in longitudinal mri studies. Neuroimage **49**(1), 391–400 (2010)
22. Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y., et al.: Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature Communications **13**(1), 6566 (2022)
23. Soh, S.E., Tint, M.T., Gluckman, P.D., Godfrey, K.M., Rifkin-Graboi, A., Chan, Y.H., Stünkel, W., Holbrook, J.D., Kwek, K., Chong, Y.S., et al.: Cohort profile: Growing up in singapore towards healthy outcomes (gusto) birth cohort study. International journal of epidemiology **43**(5), 1401–1409 (2014)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
25. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In: International Conference on Information Processing in Medical Imaging. pp. 388–400. Springer (2023)

26. Zhang, C., Adeli, E., Wu, Z., Li, G., Lin, W., Shen, D.: Infant brain development prediction with latent partial multi-view representation learning. IEEE transactions on medical imaging **38**(4), 909–918 (2018)
27. Zhang, Y., Shi, F., Cheng, J., Wang, L., Yap, P.T., Shen, D.: Longitudinally guided super-resolution of neonatal brain magnetic resonance images. IEEE transactions on cybernetics **49**(2), 662–674 (2018)