# Unisyn: A Generative Foundation Model for Universal Medical Image Synthesis across MRI, CT and PET

Yulin Wang[1], Honglin Xiong[1], Kaicong Sun[1], Jiameng Liu[1], Xin Lin[1], Ziyi Chen[1], Yuanzhe He[1], Qian Wang[1,2], Dinggang Shen[1,2,3]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
[2] Shanghai Clinical Research and Trial Center, Shanghai 201210, China
[3] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China
dgshen@shanghaitech.edu.cn

**Abstract.** Multi-modal brain imaging with MRI, CT, and PET has significantly advanced our understanding of cognition and neurodisease by providing complementary information. However, constraints on scan time and cost often result in missing critical high-quality sequences. Existing cross-modality synthesis methods are typically task- or modality-specific, leading to performance degradation when applied to heterogeneous real-world imaging data. Here, we propose UniSyn, a unified framework capable of synthesizing target imaging modalities with specific acquisition parameters from any available ones, guided by metadata. UniSyn first learns robust metadata representations through image-text alignment on large-scale multimodal neuroimaging datasets. We then introduce a cross-modality synthesis framework that leverages learned metadata representations to guide the generation of metadata-specified target images. To enhance interpretable metadata-driven control over image synthesis across diverse protocols, we design a dual-parameter arithmetic operation that explicitly integrates source and target metadata into the image translation process. Extensive experiments on multi-institutional brain imaging datasets demonstrate that UniSyn surpasses the existing cross-modality synthesis approaches in both quantitative fidelity and clinical relevance, enabling the generation of missing imaging counterparts tailored to specific clinical and research needs.

**Keywords:** Multi-modal medical image · Brain image synthesis · Vision-language model · Metadata-injecting prompt.

## 1 Introduction

Neuroimaging is essential for diagnosing and studying neurological disorders. Different modalites such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET) offer complementary insights into brain structure, function, and metabolism. For instance, MRI

provides high soft-tissue contrast without ionizing radiation, CT is optimal for detecting hemorrhages or calcifications, and PET enables metabolic assessments crucial for neurodegenerative diseases. However, acquiring multiple modalities for every patient is often impractical due to financial constraints, radiation exposure, and scanning duration limitations. The absence of a required modality can impair diagnostic accuracy and research applications, highlighting the need for a robust cross-modality synthesis framework capable of generating missing imaging counterparts.

Existing cross-modality synthesis methods predominantly focus on specific translation tasks, such as synthesizing PET from MRI [1, 2], CT from MRI [3], or inter-sequence generation [4–6], isotope tracer conversion [7], and dose modulation within the same modality [8, 9]. However, these approaches suffer from key limitations: (1) They lack generalizability across diverse imaging modalities and are constrained to fixed source-target pairs, limiting their applicability in real-world clinical scenarios. (2) Most methods operate in a purely image-driven manner, neglecting valuable non-imaging metadata such as clinical reports and imaging protocols that could enhance synthesis quality and interpretability. Previous work [10] leverages textual imaging parameters of target modality to synthesize desired MRI sequence from available ones, albeit its satisfactory synthesis performance, it is limited to inter-conversion between MRI sequences and relies on cross-attention mechanisms [11] that are sensitive to hyperparameter configurations, limiting robustness and interpretability of feature space manipulation. To address these challenges, we propose UniSyn, a unified generative framework that enables universal cross-modality brain image synthesis, guided by source and target textual metadata. UniSyn consists of two key components: (1) a metadata prompt learning network, NeuroCLIP, pre-trained on large-scale multimodal datasets, to learn a textual representation aligned with neuroimaging data, and (2) a cross-modality synthesis network, which generates anatomically and contrast-preserving target images guided by learned text embeddings. By explicitly disentangling modality-specific features from input images and integrating them with target text embeddings, UniSyn achieves improved generalization and synthesis fidelity. Extensive experiments on multi-institutional brain imaging datasets demonstrate that UniSyn outperforms existing methods in both quantitative synthesis fidelity and qualitative clinical relevance, representing a significant step toward a generalized medical image translation paradigm with broad applications in neuroimaging-based diagnosis and research.

## 2 Method

The overall framework of Unisyn is shown in Fig. 1. Specifically, we achieve text-guided image synthesis through two stages: the metadata prompt learning network (NeuroCLIP) and the generative foundation model (GFM).

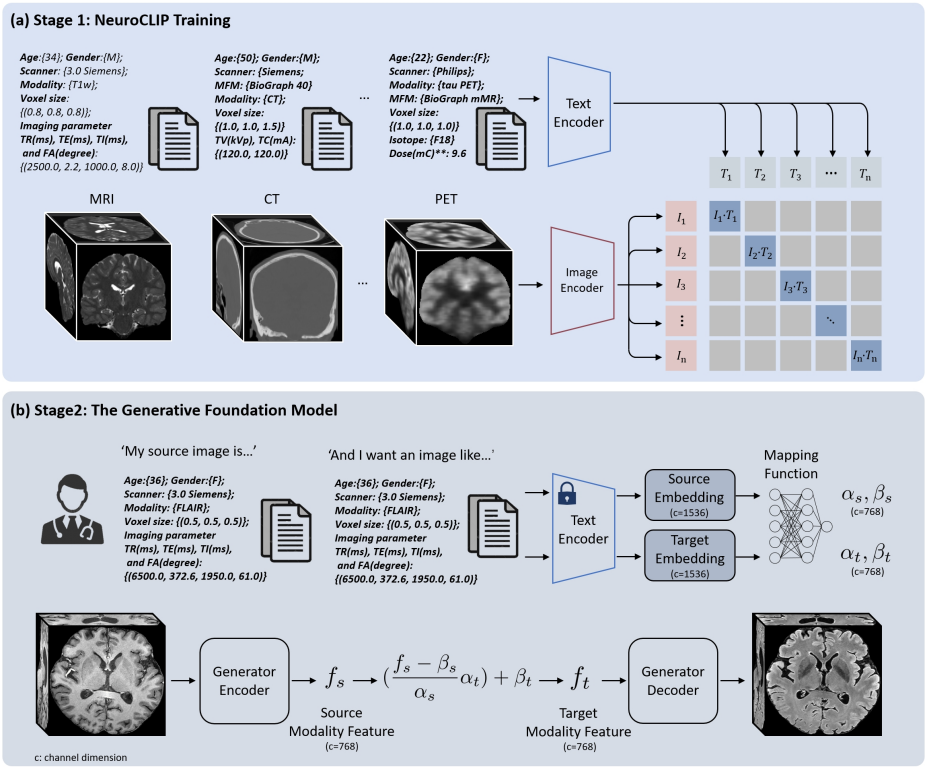


**Fig. 1.** Overview of the Unisyn framework: (a) The training of NeuroCLIP model; (b) The generative foundation model

### 2.1 NeuroCLIP

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding and processing natural language across general domains. Models like GPT-4 [12] and BERT [13] have achieved unprecedented performance in various natural language processing tasks. However, these general-purpose LLMs often struggle with domain-specific technical language, particularly in highly specialized medical fields.

Even domain-adapted models like BiomedCLIP [14], which are specifically trained on biomedical data, show limitations in comprehending specialized neuroimaging parameters such as TR (Repetition Time), TE (Echo Time), and flip angle. This gap in technical comprehension poses a significant challenge for accurate medical image synthesis and analysis.

Therefore, we propose NeuroCLIP, a specialized text encoder designed specifically for neuroimaging metadata comprehension. In Stage 1, to extract effective semantic information from textual imaging parameters and demographic attributes associated with each image, we leverage the contrastive learning strat-

egy to pretrain a text encoder that is specifically designed for 3D brain medical images' metadata understanding.

**Construction of image and text prompt dataset** The appearance of brain medical images exhibits a strong correlation with both biological attributes (e.g., age and gender) and imaging parameters. To effectively leverage these factors as guidance of GFM, we prioritize imaging parameters that most influence image intensity and contrast, and design structured textual prompts for different imaging modalities under the guidance of senior radiologists. Image samples along with their corresponding textual descriptions used for training the NeuroCLIP are presented in Fig. 1(a).

When specific imaging parameters are unavailable, "None" is used as a placeholder to maintain a consistent prompt structure. However, source and target modality information is always explicitly provided to ensure accurate modality translation.

**Contrastive learning Pre-training** To enable text representations to effectively guide cross-modality image synthesis and ensure that the generated images accurately match the specified parameters, it is essential for the text encoder to comprehensively capture the relevant imaging metadata. To achieve this, the image and text encoders are jointly trained on paired image-text samples using the contrastive learning strategy. We categorize the dataset based on imaging parameters and set the batch size accordingly. During each forward pass, a batch $B$ consists of one sample from each class, and the embedding distance between paired image-text samples is minimized, while the distance between non-paired samples is maximized by computing contrastive loss.

The model architecture is built upon [10], with the encoded token length extended from 90 to 224, thereby improving the text encoder's ability to process longer text prompts and support diverse downstream tasks.

Upon completion of the Stage 1 training, the pretrained text encoder is frozen and subsequently utilized in Stage 2.

### 2.2 The generative foundation model

Stage 2 focuses on generating the target image $I_t$ from any available source image $I_s$ by leveraging both source $T_s$ and target text prompts $T_t$ as guidance. As illustrated in Fig.1(b), the synthesis process begins by encoding $I_s$, $T_s$, and $T_t$ into corresponding feature embeddings, $T_s$ and $T_t$ are then used to guide the image feature transformation process to obtain the target image feature $f_t$. Finally, $f_t$ is reconstructed to the target image $\hat{I}_t$ by the image decoder.

**Text-guided Representation Learning** The representation learning process begins with encoding source text prompts into feature vectors. Inspired by [15], we decompose medical image volumes into two fundamental components: content

features that capture anatomical structures, and contrast features that encode modality-specific characteristics.

To effectively model the modality-specific contrast characteristics, we introduce a novel mapping mechanism that transforms text features into two key parameters: a scaling factor ($\alpha$) and a bias term ($\beta$). Specifically, the text prompt is first encoded into a 1536-dimensional feature representation, where the first 768 dimensions are assigned to $\alpha$ and the remaining 768 dimensions to $\beta$. These parameters are designed to capture the essential aspects of cross-modality variations: $\alpha$ models contrast differences between imaging modalities, while $\beta$ accounts for baseline intensity shifts arising from diverse acquisition protocols.

Our framework implements a systematic transformation pipeline for cross-modality synthesis: First, we normalize the source image feature $f_s$ (768 dimensions) to obtain $f_0$ by applying: $f_0 = (f_s - \beta_s)/\alpha_s$. This normalization step isolates the modality-independent anatomical representation. Subsequently, we synthesize the target feature $f_t$ by incorporating the target modality characteristics: $f_t = \alpha_t f_0 + \beta_t$, where $\alpha_t$ and $\beta_t$ are the scaling and bias parameters specific to the target modality. Finally, the transformed feature $f_t$ is processed through a decoder to generate the synthesized target image $\hat{I}_t$. This formulation enables explicit and interpretable modulation of the image feature space, facilitating effective cross-modality synthesis while preserving anatomical integrity.

### Model Architecture

*Image Encoder* : The image encoder consists of 24 residual blocks (ResBlocks), with residual connections between each block to facilitate gradient propagation. Each ResBlock consists of two sequential $3 \times 3$ convolutional layers, followed by a ReLU activation layer. The encoder is designed with a channel dimension of 1024, ensuring sufficient capacity for feature extraction.

*Image Decoder* : The decoder is a four-layer fully connected network with ReLU activation. The dimensions of each layer are 2048, 1024, 1024, and 8, respectively. The final layer with 8 output channels matches the input channel number of the image encoder, facilitating the image synthesis from processed feature maps.

**Loss Functions** To strengthen the constraint on general feature representation learning, we incorporate a cyclic synthesis strategy during training. Specifically, within each iteration, the source image and target image are utilized to generate each other, and their encoded features $F_x$ and $F_y$ are enforced to be consistent. We impose a pixel-level similarity as the supervisory signal.

$$\mathcal{L}_f = \sum_i \|F_x - F_y\|_1 \tag{1}$$

We also introduce a synthesis loss as a supervisory signal for cross-contrast translation by directly comparing the synthesized images with their corresponding

ground truth counterparts. The synthesis loss is then computed as the L1-norm of the residual error:

$$\mathcal{L}_s = \sum_i \|\hat{X}_i - X_i\|_1 \tag{2}$$

Combining the two loss terms above, we derive the overall loss as:

$$\mathcal{L}_{total} = \lambda_f \mathcal{L}_f + \lambda_s \mathcal{L}_s \tag{3}$$

## 3 Experiments and Results

### 3.1 Datasets and Data Preprocessing

**Description of the GFM Dataset**

*Huashan Dataset.* The Huashan (HS) dataset consists of 1,002 participants for the Alzheimer's Disease Study, recruited from both the Universal Medical clinic and Shanghai Sixth People's Hospital, China. From this cohort, we selected a subset of 847 individuals who all underwent T1-weighted (T1w) MRI scans. Among these participants, the availability of additional imaging modalities—CT scans, amyloid-beta (A$\beta$)-PET scans, Fluorodeoxyglucose (FDG)-PET scans, and other MRI sequences, including T2-weighted (T2w) and Fluid-Attenuated Inversion Recovery (FLAIR)—varied on a per-subject basis.

*Zhongshan Dataset.* The Zhongshan (ZS) dataset comprises 114 participants from Zhongshan Hospital, Fudan University, enrolled for the neuro-degradation study. Each subject includes T1w, T2w, and FLAIR MRI sequences, CT, and FDG-PET scans acquired using United Imaging Healthcare (UIH) scanning systems.

**Description of the NeuroCLIP Dataset.** For NeuroCLIP pretraining, except for the above two datasets, we use a large-scale dataset containing 47,841 3D scans collected from multiple centers, covering various imaging parameters. These datasets include T1w, T2w, FLAIR, PD, SWI, T2Star, T1CE MRI sequences, PET imaging with FDG, AV45, and TAU deposition, and CT.

For image preprocessing, all imaging modalities are spatially registered to the corresponding subject's T1w image, preserving its original spatial resolution. Subsequently, skull stripping is performed on all images. The dataset is then split into training, validation, and test sets using a 7:1:2 ratio.

### 3.2 Implementation Details

In our experiments, both the NeuroCLIP and GFM are trained on a single NVIDIA A100 GPU with 40GB memory. For NeuroCLIP, both image and text encoders are trained from scratch for 100 epochs using the Adam optimizer [16]. For the GFM, its image encoder and decoder are trained from scratch, with an

input and output patch size of $8 \times 64 \times 64$. GFM is trained for 300 epochs using the Adam optimizer, with a batch size of 8, an initial learning rate of $10^{-4}$, and a weight decay by 0.5 every 100 epochs. We empirically set $\lambda_f = 0.1$, $\lambda_s = 1.0$ in our experiments. To comprehensively evaluate model performance on image synthesis, we employed Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as quantitative metrics.

### 3.3 Comparison with State-of-the-art Methods

We conducted comprehensive experiments on the HS and ZS datasets to evaluate the effectiveness of our framework against three competitive state-of-the-art (SOTA) medical image synthesis models: 1) PTNet [5], an advanced 3D transformer-based model for infant brain MRI sequences synthesis; 2) BME-X [17], a unified model for the motion correction, resolution enhancement, denoising and harmonization of MR images; and 3) TUMSyn [10], which utilizes textual metadata as prompt for any MRI sequences synthesis. All methods are in their default configuration.

**Quantitative Comparison Analysis.** Table 1 presents the quantitative evaluation of all methods across six clinically significant tasks. For PET image synthesis tasks, we synthesized FDG-PET images for the ZS dataset and AV45-PET images for the HS dataset. Our proposed framework consistently achieves the highest PSNR and SSIM scores across all tasks, outperforming three SOTA methods. This performance enhancement highlights that explicit integration of both source and target textual metadata, along with the joint training strategy, can collectively empower the model to capture robust shared features across varying imaging parameters, thereby mitigating noise and artifacts introduced during cross-modality transformations.

**Qualitative Comparison Analysis.** To qualitatively assess the effectiveness of our proposed framework, we conducted visual comparisons across four representative tasks: T1w-to-CT, T1w-to-PET, CT-to-PET, and T1w-to-FLAIR. As

**Table 1.** Quantitative comparison across different tasks.

| | | ZS T1-CT | ZS CT-PET | ZS T1-PET | HS T1-CT | HS T1-FLAIR | HS T1-PET |
|---|---|---|---|---|---|---|---|
| PTNet | PSNR | 25.78±1.19 | 24.63±0.38 | 25.89±0.51 | 24.91±2.22 | 27.21±3.01 | 23.89±2.32 |
| | SSIM | 0.918±0.033 | 0.885±0.020 | 0.906±0.027 | 0.905±0.027 | 0.940±0.048 | 0.889±0.034 |
| BME-X | PSNR | 24.81±1.21 | 23.56±0.51 | 24.71±0.64 | 23.85±2.44 | 25.83±3.13 | 22.53±2.37 |
| | SSIM | 0.902±0.030 | 0.872±0.026 | 0.896±0.029 | 0.894±0.040 | 0.933±0.049 | 0.871±0.037 |
| TUMSyn | PSNR | 27.56±1.29 | 24.90±0.49 | 26.98±0.43 | 26.33±2.28 | 28.14±3.07 | 24.78±2.18 |
| | SSIM | 0.930±0.034 | 0.893±0.021 | 0.913±0.027 | 0.915±0.033 | 0.958±0.042 | 0.894±0.028 |
| Unisyn | PSNR | **28.04±1.36** | **25.75±0.53** | **27.45±0.79** | **26.78±2.32** | **29.05±2.99** | **25.47±2.14** |
| | SSIM | **0.941±0.034** | **0.909±0.028** | **0.926±0.032** | **0.927±0.035** | **0.965±0.039** | **0.907±0.027** |

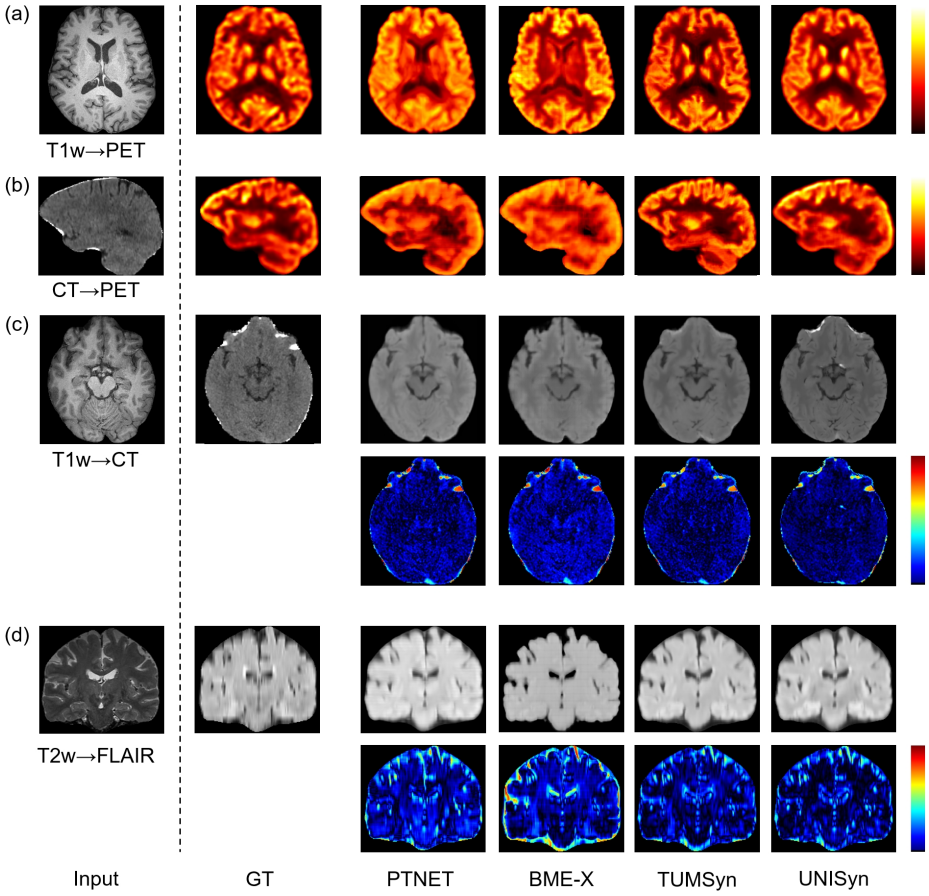


**Fig. 2.** Results of four cross-modality synthesis tasks from different methods.

illustrated in Fig. 2, the synthesized images generated by UniSyn closely resemble the ground truth (GT), demonstrating superior anatomical detail preservation and contrast consistency compared to three SOTA methods. These findings align with the quantitative evaluation results. The same conclusion is provided by the error maps (second row of each task), which highlight the reduced discrepancies in UniSyn's outputs.

### 3.4 Ablation Study

To substantiate the effectiveness of our text-guided representation learning and the jointly training strategy, we conduct ablation experiments by replacing the baseline task-specific model, which employs a cross-attention mechanism, with our proposed arithmetic operation. Additionally, we compare the unified training strategy that incorporates all imaging modalities and tasks.

**Table 2.** Comparison of baseline and Unisyn models

| | Baseline | | | Unisyn (specific task) | | | Unisyn | |
|---|---|---|---|---|---|---|---|---|
| **Task** | **PSNR** | **SSIM** | **Inference time (s)** | **PSNR** | **SSIM** | **Inference time (s)** | **PSNR** | **SSIM** |
| CT-T1w | 25.53±2.09 | 0.879±0.018 | 72.8 | 26.01±2.03 | 0.893±0.018 | 50.2 | **26.54±2.19** | **0.901±0.021** |
| T1w-T2w | 29.39±2.16 | 0.938±0.025 | 81.1 | 30.43±2.17 | 0.951±0.027 | 58.5 | **30.74±2.24** | **0.958±0.028** |

As detailed in Table 2, the baseline method, which relies on cross-attention mechanisms for target text prompt and source image fusion (Baseline), our arithmetic operation (Unisyn (specific task)) achieves mean improvements of 1.9% in PSNR and 1.6% in SSIM across both tasks. Furthermore, applying the joint training strategy further enhances 0.4 dB PSNR and 0.008 SSIM.

These findings indicate that employing arithmetic operations for image-text fusion significantly reduces the model's learning burden regarding complex interactions between these modalities. Concurrently, the integration of textual metadata facilitates unified training across heterogeneous imaging modalities, subsequently enhancing the robustness of synthetic image generation.

## 4 Conclusion

In this study, we propose UniSyn, a universal framework for the synthesis of diverse neuroimaging modalities from any available ones, guided by textual metadata. By initially pre-training a domain-specific text encoder, UniSyn effectively captures the factors that differentially influence image contrast and anatomical content from metadata, enabling unified, parameter-customized image synthesis. To seamlessly integrate imaging and non-imaging data, we introduce an arithmetic operation that enhances the precision and robustness of source-target image feature mapping. Extensive experiments on heterogeneous datasets demonstrate UniSyn's ability to generate high-fidelity 3D brain MRI, PET, and CT volumes, highlighting its potential to address the challenge of missing multimodal neuroimaging data in clinical and research settings.

**Acknowledgments.** This study was funded in part by National Natural Science Foundation of China (grant numbers 82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (S20240085, STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), Shanghai Municipal Central Guided Local Science and Technology Development Fund (No. YDZX20233100001001), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), and HPC Platform of ShanghaiTech University.

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

## References

1. Hussein, R., Shin, D., Zhao, M.Y., Guo, J., Davidzon, G., Steinberg, G., Moseley, M., Zaharchuk, G.: Turning brain mri into diagnostic pet: 15o-water pet cbf

synthesis from multi-contrast mri via attention-based encoder–decoder networks. Medical Image Analysis **93**, 103072 (2024)
2. Ou, Z., Pan, Y., Xie, F., Guo, Q., Shen, D.: Image-and-label conditioning latent diffusion model: Synthesizing a$\beta$-pet from mri for detecting amyloid status. IEEE Journal of Biomedical and Health Informatics (2024)
3. Pan, S., Abouei, E., Wynne, J., Chang, C.W., Wang, T., Qiu, R.L., Li, Y., Peng, J., Roper, J., Patel, P., et al.: Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. Medical Physics **51**(4), 2538–2548 (2024)
4. Wang, Y., Wu, W., Yang, Y., Hu, H., Yu, S., Dong, X., Chen, F., Liu, Q.: Deep learning-based 3d mri contrast-enhanced synthesis from a 2d noncontrast t2flair sequence. Medical Physics **49**(7), 4478–4493 (2022)
5. Zhang, X., He, X., Guo, J., Ettehadi, N., Aw, N., Semanek, D., Posner, J., Laine, A., Wang, Y.: Ptnet3d: A 3d high-resolution longitudinal infant brain mri synthesizer based on transformers. IEEE transactions on medical imaging **41**(10), 2925–2940 (2022)
6. Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. IEEE Transactions on Medical Imaging (2023)
7. Zhou, B., Wang, R., Chen, M.K., Mecca, A.P., O'Dell, R.S., Van Dyck, C.H., Carson, R.E., Duncan, J.S., Liu, C.: Synthesizing multi-tracer pet images for alzheimer's disease patients using a 3d unified anatomy-aware cyclic adversarial network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 34–43. Springer (2021)
8. Xue, Y., Bi, L., Peng, Y., Fulham, M., Feng, D.D., Kim, J.: Pet synthesis via self-supervised adaptive residual estimation generative adversarial network. IEEE Transactions on Radiation and Plasma Medical Sciences (2023)
9. Jiang, C., Pan, Y., Cui, Z., Nie, D., Shen, D.: Semi-supervised standard-dose pet image generation via region-adaptive normalization and structural consistency constraint. IEEE transactions on medical imaging **42**(10), 2974–2987 (2023)
10. Wang, Y., Xiong, H., Sun, K., Bai, S., Dai, L., Ding, Z., Liu, J., Wang, Q., Liu, Q., Shen, D.: Toward general text-guided multimodal brain mri synthesis for diagnosis and medical image analysis. Cell Reports Medicine (2025)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
12. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
14. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
15. Xiong, H., Fang, Y., Sun, K., Wang, Y., Zong, X., Zhang, W., Wang, Q.: Contrast representation learning from imaging parameters for magnetic resonance

image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 181–190. Springer (2024)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nature Biomedical Engineering pp. 1–18 (2024)