This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Determinism, Plausibility, and Clinical Trust in CT-to-MRI Translation for Emergency Spine Imaging

Paul Banahan[1,2,3,4], Prateek Mathur[1,2], Molly Godson Treacy[3], Edward McDermott[3], Eimear Kyle[3], David Caldwell[3], Aonghus Lawlor[1,2], and Peter MacMahon[3]

[1] School of Computer Science, University College Dublin, Ireland
[2] Rinn Artificial Intelligence: Research & Innovation in Data Science and AI
[3] Mater Misericordiae University Hospital
[4] Pillar Centre for Transformative Healthcare

Abstract. In suspected Cauda Equina Syndrome (CES), MRI is the reference standard for diagnosis and surgical decision-making, but access may be delayed or contraindicated in emergency care. CT is more readily available, motivating interest in CT-to-MRI translation as a potential clinical support tool. However, in high-risk settings, synthetic images must be evaluated beyond visual similarity alone. They must preserve clinically relevant anatomy and pathology, and remain consistent given identical generation conditions. In this pilot study, we evaluate five representative CT-to-MRI translation paradigms for emergency spine imaging through three axes: plausibility, clinical trust, and determinism. Plausibility is assessed using conventional quantitative metrics and semantic representation-level analysis. Clinical trust is assessed through blinded radiologist evaluation of image quality, anatomy, pathology, and confidence. Determinism is assessed using repeated-generation coefficient-of-variation analysis. Our findings show that favourable quantitative performance does not consistently translate into stronger clinician-derived assessment, and that repeat-generation stability provides a separate reliability indication. These results support multi-dimensional, clinically informed evaluation of generative medical imaging systems before further clinical translation.

Keywords: Generative AI · Clinical Evaluation · Image-to-Image Translation · Clinical Reliability

1 Introduction

In emergent spinal pathologies such as Cauda Equina Syndrome (CES), magnetic resonance imaging (MRI) is the reference standard for diagnosis and surgical decision-making [13]. CES involves compression of the cauda equina nerves, critical for lower limb and pelvic function. However, MRI access may be delayed or contraindicated due to resource constraints or patient factors such as

non–MRI-safe implants or claustrophobia [2]. Computed tomography (CT) is widely available and rapid in acute settings [5], but provides limited soft-tissue contrast for evaluating spinal pathology. This gap between imaging availability and diagnostic standards has motivated interest in CT-to-MRI translation, where generative models synthesise MRI-like images from routinely available CT scans [10,12].

Generative models, including adversarial and diffusion-based approaches, have shown promise for cross-modal translation in medical imaging [8,16]. Conditioning strategies based on anatomical annotations, soft-tissue priors or paired source-target information may further improve structural consistency in image-to-image translation [7,12,11]. However, in emergency imaging, the clinical value of a synthetic image depends on more than just visual realism. A generated image must preserve relevant anatomy, avoid suppressing subtle pathology and behave consistently across repeated use.

Evaluation of synthetic images has therefore become a central challenge. Studies have shown that commonly used similarity measures, such as the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), can underestimate blurring and correlate poorly with human perception [3,4]. Furthermore, favourable numerical performance does not guarantee faithful pathology modelling [12]. In emergency imaging, particularly in surgical decision-making, the key question is not whether synthetic images resemble real MRI, but whether they preserve diagnostically actionable information and behave consistently. The relationship between quantitative fidelity, perceptual plausibility, and clinical reliability therefore remains underexplored.

In this pilot study, we evaluate five representative CT-to-MRI translation paradigms for emergency spine imaging across three axes: determinism, plausibility, and clinical trust. Plausibility is assessed using quantitative similarity metrics, representation-level feature analysis, and radiologist-rated image quality. Clinical trust is assessed using radiologist judgements of relevant anatomy, pathology, and associated confidence. Determinism is assessed through repeated generation under identical conditioning and coefficient-of-variation analysis. This framework allows us to examine whether conventional measures of image plausibility align with clinically grounded and stability-based measures of reliability.

2 Methods

2.1 Dataset

The dataset consists of 66 patients with paired lumbar spine CT and T2-weighted MRI scans acquired within 24 hours of each other for patients presenting to the emergency department with suspected CES. CT images were acquired using an optimised protocol designed to improve CES detection and soft-tissue conspicuity [2]. All image pairs have been co-registered by expert radiologists, with annotations available for multiple anatomical structures (e.g., vertebrae, aorta, and organs). Patients were split into training, validation, and test sets using a

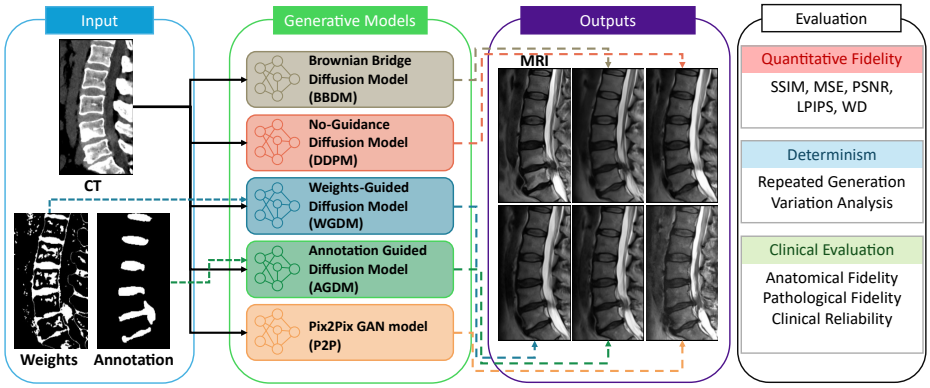


Fig. 1: Overview of the CT-to-MRI translation evaluation framework. Five representative generative paradigms with varying levels of conditioning are tested. Synthetic MRI outputs are assessed using quantitative fidelity metrics, repeated-generation variability analysis, and blinded clinical assessment.

70:20:10 patient-level split. Since the target MRI was acquired as sagittal 2D T2-weighted imaging, translation was performed slice-wise in the sagittal plane to match the native acquisition geometry. Images were cropped from the L1 vertebra to the sacrum, resized to 256×256 pixels, and min–max normalised before model input.

2.2 Model Selection

Figure 1 summarises the study design. We selected five representative paradigms for paired 2D CT-to-MRI translation, spanning adversarial and diffusion-based synthesis with varying levels of conditioning and stochasticity. All models were

No.	Question	Scale
Q1	How would you rate the overall image quality of this sequence?	1-10
Q2	How would rate the consistency of the image slices?	1-10
Q3	Is the image sequence noisy or overprocessed/overly smooth?	1-10
Q4	In your assessment is the vertebral canal clear?	Yes/No
Q5	How confident are your answer to the above question?	1-10
Q6	Do you see an emergency pathology or abnormality like cauda equina, fractures or a tumour?	Yes/No
Q7	How confident are your answer to the above question?	1-10

Table 1: Survey Questions and Rankings[1]

trained using identical patient splits and preprocessing to support direct comparison.

Pix2Pix [7] was included as a deterministic adversarial benchmark for cross-modal image translation in medical imaging [14,12]. Three diffusion-based variants [6] were evaluated: a CT-conditioned DDPM baseline, an annotation-guided diffusion model (AGDM) conditioned on expert anatomical annotations, and a weight-guided diffusion model (WGDM) conditioned on CT-derived soft-tissue weight maps. The soft-tissue maps were generated by applying a soft-tissue CT window (level 125, width 150), followed by intensity normalisation, high-intensity thresholding to suppress osseous structures, and re-normalisation of the residual image to emphasise soft-tissue regions. Finally, a Brownian Bridge Diffusion Model (BBDM) [9] was included to represent bridge-based conditional diffusion, where generation is formulated as a stochastic transition between the CT and MRI domains.

2.3 Quantitative and Representation-Level Evaluation

Quantitative fidelity was assessed using conventional image-similarity metrics [3]. We included MSE, PSNR, and SSIM to capture pixel-wise error, signal-to-error fidelity, and structural similarity. To complement these measures, we also included Wasserstein distance (WD), which estimates distributional differences between real and synthetic image intensities, and learned perceptual image patch similarity (LPIPS), which measures perceptual similarity using deep feature representations.

To further examine whether image-level metrics were supported by higher-level feature agreement, we performed an exploratory representation-level analysis using embeddings extracted from the pre-trained CLIP vision encoder [15]. Cosine similarity between synthetic and real MRI embeddings was computed as a feature-space proximity measure. To visualise cross-domain organisation, embeddings from CT, real MRI, and synthetic images were projected into two dimensions using t-distributed stochastic neighbour embedding (t-SNE).

All slice-wise quantitative measures were evaluated at the patient level to avoid pseudo-replication. For each model and metric, slice-wise values were first averaged within each patient, yielding one observation per patient-model pair. Inter-model differences were assessed using the non-parametric repeated-measures Friedman test. Where appropriate, post-hoc paired comparisons were performed using the Wilcoxon signed-rank test. Given the limited test cohort, these analyses were interpreted as exploratory rather than as definitive evidence of model superiority.

2.4 Clinician Evaluation

Clinical reader scores were obtained from our previously conducted blinded and randomised reader assessment study [1]. Five expert radiologists were shown cine clips of sagittal image sequences from real and synthetic images and completed a structured seven-question survey designed with consultant radiologists to ensure

clarity while preserving radiological intent (see Table 1). The survey included the held-out test set of five patients, four of whom presented with emergent spinal pathologies including CES, fractures or tumours.

Survey responses were distilled into three components: anatomical fidelity, pathological fidelity, and image fidelity. Question 1 (overall image quality) was retained as an independent perceptual ranking metric. Image fidelity was derived from inter-slice consistency (Q2) and perceived noise (low score) or over-smoothing (high score) (Q3). Since Q3 penalises both excessive noise and smoothing, ratings were re-centred to favour mid-scale responses. Let $Q_3 \in \{1, \ldots, 10\}$ denote the original score and define

$$Q_3^* = 10 - |Q_3 - 5|, \qquad S_{\text{imag}} = \frac{Q_2 + Q_3^*}{2}, \tag{1}$$

so that values closer to the midpoint correspond to higher perceived quality and S_{imag} represents the averaged image fidelity score.

Anatomical fidelity was defined using vertebral canal assessment (Q4) and associated confidence (Q5), based on concordance with the reader's response for the ground-truth MRI. Let A_{MRI} and A_{syn} denote binary responses for real and synthetic images, and $Q_5 \in \{1, \ldots, 10\}$ the diagnostic confidence. Pathological fidelity was computed analogously using pathology detection (Q6) and confidence (Q7), with P_{MRI} and P_{syn} denoting pathology presence and Q_7 the associated confidence. The derived scores were defined as:

$$S_{\text{anat}} = \begin{cases} 5 + \frac{Q_5}{2}, & \text{if } A_{\text{MRI}} = A_{\text{syn}}, \\ 5 - \frac{Q_5}{2}, & \text{otherwise,} \end{cases} \qquad S_{\text{path}} = \begin{cases} 5 + \frac{Q_7}{2}, & \text{if } P_{\text{MRI}} = P_{\text{syn}}, \\ 5 - \frac{Q_7}{2}, & \text{otherwise.} \end{cases} \tag{2}$$

2.5 Determinism Analysis

Determinism, defined here as stable generation given identical input conditions, is an important pre-requisite for clinical trust. Variation in synthetic anatomy or pathology could undermine reader confidence in the synthetic output. We

Table 2: Patient-level quantitative similarity metrics (mean ± standard deviation) for CT-to-MRI translation models. Slice-wise values were first averaged per patient to avoid pseudo-replication. Arrows indicate the direction of optimal performance (↑ higher is better; ↓ lower is better). Best-performing value per metric is highlighted in bold.

Method	MSE ↓	SSIM ↑	PSNR ↑	WD ↓	LPIPS ↓
Pix2Pix	101.201 ± 2.386	0.389 ± 0.056	28.083 ± 0.101	21.313 ± 5.807	0.400 ± 0.021
DDPM	97.594 ± 1.513	0.443 ± 0.059	28.247 ± 0.072	19.374 ± 4.174	0.404 ± 0.059
WGDM	97.337 ± 1.831	0.433 ± 0.061	28.257 ± 0.087	17.294 ± 4.670	0.400 ± 0.056
AGDM	97.615 ± 1.206	0.439 ± 0.060	28.243 ± 0.057	18.514 ± 4.436	**0.398 ± 0.057**
BBDM	**94.059 ± 5.016**	**0.445 ± 0.064**	**28.410 ± 0.242**	**16.048 ± 7.561**	0.424 ± 0.058

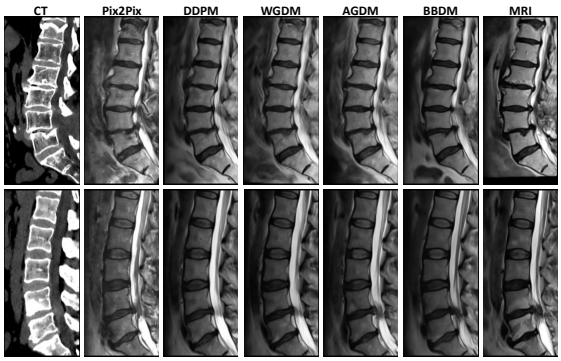
assessed determinism by a coefficient of variation analysis. Given each input (i) and fixed model weights, sampling is performed by varying the stochastic seed (r) and computing a coefficient of variation (CV) defined as:

$$CV_i = \frac{\mathrm{std}_r\left(\langle \hat{x}_i^{(r)} \rangle_\Omega\right)}{\mathrm{mean}_r\left(\langle \hat{x}_i^{(r)} \rangle_\Omega\right)}, \tag{3}$$

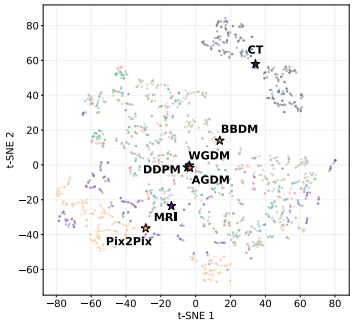
where $(\langle \cdot \rangle_\Omega)$ denotes the spatial mean over the image domain. In this work, five repeated generations were computed for each input.

3 Results and Discussions

Plausibility. Figure 2(a) presents qualitative examples of generated images given CT and target MRI. The generated images reproduced the broad sagittal spine appearance of the target but differed in soft-tissue features, smoothness and anatomical detail across models. Table 2 summarises patient-level quantitative fidelity metrics. BBDM achieved the strongest mean performance for MSE, SSIM, PSNR and WD, suggesting favourable quantitative plausibility under conventional image-similarity metrics. However, BBDM also showed greater inter-patient variability, indicating that this average may be case-dependent. Friedman testing showed detectable inter-model differences for MSE, SSIM, and PSNR, but not for WD, while LPIPS showed borderline separation. This suggests that model differences were more apparent under conventional pixel-level

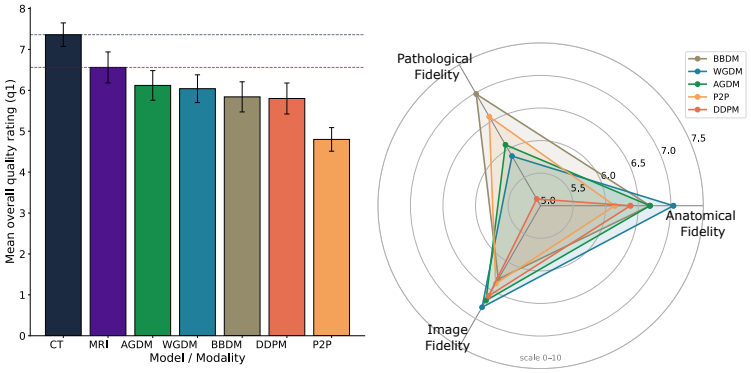


(a) Qualitative Comparison of generated samples



(b) t-SNE decomposition of CLIP embeddings of generated images

Fig. 2: Qualitative and representation-level comparison of CT-to-MRI translation models. (a) Representative sagittal slices showing CT input, synthetic MRI outputs from each model, and the corresponding ground-truth MRI. (b) t-SNE projection of CLIP feature embeddings for CT, real MRI, and generated images. Model centroids are indicated.



(a) Average ranking for perceived quality (Q1).

(b) Average derived scores across evaluation domains (Q2–Q7).

Fig. 3: Average rankings for perceived image quality (a) and aggregated derived scores (b) showing the mean ranking assigned to perceived quality and the corresponding average derived scores across the remaining evaluation domains.

and structural fidelity measures than under distributional or perceptual measures. Post-hoc comparisons did not establish robust pairwise superiority, indicating that the quantitative results should be interpreted as metric-dependent trends rather than definitive evidence of model dominance.

The representation-level analysis in Figure 2(b) provides a complementary view of plausibility. The t-SNE projection shows that model families occupy distinct regions between the CT and MRI domains, suggesting that different generative paradigms produce MRI-like images with different degrees of feature-space alignment. Together with the quantitative results, this supports the interpretation that plausibility is not captured by a single similarity metric, but varies across pixel-level, perceptual and representation-level assessments.

Clinical Trust. Although BBDM achieved the strongest mean quantitative performance, this superiority was not observed in the clinician evaluation. As shown in Figure 3(a), the ground-truth MRI received the highest perceived quality rating despite blinding, reinforcing its position as the clinical standard. Among synthetic outputs, guided diffusion models (AGDM and WGDM) received better perceived quality ratings overall followed by BBDM and DDPM, whereas Pix2Pix was consistently rated low. The derived scores in Figure 3(b) further demonstrate that reader assessment varied across clinical domains. BBDM and Pix2Pix showed comparatively stronger pathological than anatomical fidelity whereas WGDM showed stronger anatomical and moderate pathological fidelity. Importantly, no model consistently dominated across the evaluation axes. The feature-space separation observed in Figure 2(b) is consistent with the image fidelity axis of the reader evaluation. However, it also highlights that

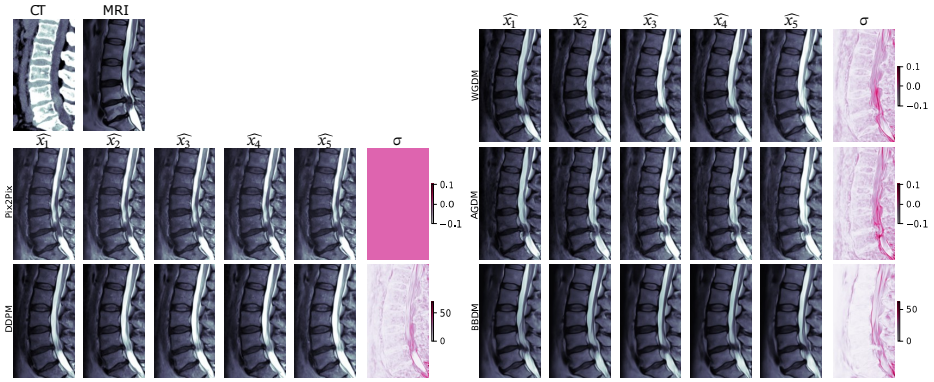


Fig. 4: Determinism analysis under repeated sampling. Five independent generations per input are shown for each model ($\hat{x}_1$–$\hat{x}_5$), with the corresponding pixel-wise standard deviation map (σ).

clinical judgement for anatomical and pathological fidelity is not consistently captured by either aggregate similarity metrics or representation-level analysis.

Determinism. Figure 4 illustrates the sample examples under repeated generations given identical inputs. As can be observed, Pix2Pix produced near-identical outputs across runs, consistent with deterministic inference (CV ≈ 0). Among diffusion-based models, BBDM exhibited the lowest variability (CV = 0.0207), followed by WGDM (CV = 0.0987), DDPM (CV = 0.117), and AGDM (CV = 0.1211). Variance maps indicate that stochastic differences are concentrated around anatomical boundaries and pathology-relevant regions. Determinism adds a separate reliability check by asking whether the same clinical input produces a stable synthetic representation. For stochastic generative models, this matters because visually plausible outputs may still vary across repeated sampling. In emergency imaging, variability near anatomical boundaries or pathology-relevant regions could affect reader confidence, even when average image fidelity appears favourable.

4 Conclusions

In this pilot study, we examined CT-to-MRI translation for emergency spine imaging through three axes: determinism, plausibility and clinical trust. Our findings suggest that clinical reliability cannot be inferred from quantitative image-similarity metrics alone. Models with favourable fidelity scores did not consistently dominate in the clinician-derived assessments, and repeat-generation stability provided a separate reliability indication. These results support multi-dimensional validation of synthetic medical images, particularly in emergency settings where visual plausibility, clinical interpretability and stable generation are all required.

This study is limited by a single-centre cohort, the use of general-purpose CLIP feature representations and clinician assessment through a structured survey rather than a standard radiology workflow. Future work will expand evaluation to more datasets and use clinically aligned reader studies.

Prospects of Application. CT-to-MRI translation could address an unmet clinical need when emergency MRI is delayed, unavailable, or contraindicated for CES. However, clinical translation must require validation beyond quantitative fidelity, including image plausibility, repeat-generation determinism, and clinician-centred trust in anatomical and pathological representation.

Ethics Approval Approval was granted by the Mater Misericordiae University Hospital's Research Ethics Committee. All data used in this study was fully and irreversibly anonymised.

Acknowledgments. This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland, under Grant number [23/RC/13506] at the Research Ireland Centre - Rinn Artificial Intelligence. Additionally, this project received grants from the Mater Hospital Foundation and the UCD AI Healthcare Hub.

Disclosure of Interests. No competing interests to declare.

References

1. Banahan, P., McDermott, E., Treacy, M.G., Mathur, P., Caldwell, D., Kyle, E., Lawlor, A., MacMahon, P.: Quantitative and qualitative assessment of generative AI for cross-modal CT-to-MRI translation in spine imaging. 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI) **00**, 1–4 (2026). https://doi.org/10.1109/isbi61048.2026.11516047
2. Dempsey, P.J., Ryan, D.T., Lambe, G., Power, J.W., Yates, A.H., Kenny, G., MacMahon, P.J.: An optimized ct protocol for detecting suspected cauda equina syndrome: A comparative analysis with mri. American Journal of Neuroradiology **45**(9), 1370–1377 (Jul 2024). https://doi.org/10.3174/ajnr.a8315, http://dx.doi.org/10.3174/ajnr.A8315
3. Dohmen, M., Klemens, M.A., Baltruschat, I.M., Truong, T., Lenga, M.: Similarity and quality metrics for MR image-to-image translation. Scientific Reports **15**(1), 3853 (2025). https://doi.org/10.1038/s41598-025-87358-0
4. Dohmen, M., Truong, T., Baltruschat, I.M., Lenga, M.: Deep generative models, 4th MICCAI workshop, DGM4miccai 2024, held in conjunction with MICCAI 2024, marrakesh, morocco, october 10, 2024, proceedings. Lecture Notes in Computer Science pp. 150–159 (2024). https://doi.org/10.1007/978-3-031-72744-3_15
5. Ginde, A.A., Foianini, A., Renner, D.M., Valley, M., Camargo, Jr, C.A.: Availability and quality of computed tomography and magnetic resonance imaging equipment in us emergency departments. Academic emergency medicine **15**(8), 780–783 (2008)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv (2020). https://doi.org/10.48550/arxiv.2006.11239

7. 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 (CVPR) (July 2017)
8. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models for medical image analysis: A comprehensive survey. arXiv (2022). https://doi.org/10.48550/arxiv.2211.07804
9. Li, B., Xue, K., Liu, B., Lai, Y.K.: BBDM: Image-to-image translation with brownian bridge diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) **00**, 1952–1961 (2023). https://doi.org/10.1109/cvpr52729.2023.00194
10. Li, H.B., et al.: The brain tumor segmentation (BraTS) challenge 2023: Brain MR image synthesis for tumor segmentation (BraSyn). ArXiv (2024)
11. Luo, Y., Zhang, S., Ling, J., Lin, Z., Wang, Z., Yao, S.: Mask-guided generative adversarial network for MRI-based CT synthesis. Knowledge-Based Systems **295**, 111799 (2024). https://doi.org/10.1016/j.knosys.2024.111799
12. Mathur, P., Banahan, P., Burns, J., MacMahon, P., Lawlor, A.: Simulation and synthesis in medical imaging, 10th international workshop, SASHIMI 2025, held in conjunction with MICCAI 2025, daejeon, south korea, september 23, 2025, proceedings. Lecture Notes in Computer Science pp. 139–148 (2025). https://doi.org/10.1007/978-3-032-05573-6_14
13. McNamee, J., Flynn, P., O'Leary, S., Love, M., Kelly, B.: Imaging in cauda equina syndrome–a pictorial review. The Ulster Medical Journal **82**(2), 100 (2013)
14. Moschetto, A., Puglisi, L., Sargood, A., Dell'Acqua, P., Guarnera, F., Battiato, S., Ravì, D.: Benchmarking GANs, diffusion models, and flow matching for t1w-to-t2w MRI translation. arXiv (2025). https://doi.org/10.48550/arxiv.2507.14575
15. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
16. Skandarani, Y., Jodoin, P.M., Lalande, A.: GANs for medical image synthesis: An empirical study. arXiv (2021). https://doi.org/10.48550/arxiv.2105.05318