Time Matters: Rethinking Diffusion and Flow Models in One-Step Medical Image Translation

Siyuan Mei[1(✉)], Yanteng Zhang[2(✉)], Yan Xia[3], Qizhen Lan[4], Yipeng Sun[1], Siming Bayer[1], Zirong Li[3], Chengze Ye[1], Daiqi Liu[1], Xiaoqian Jiang[4], Fuxin Fan[5], Yixing Huang[6], and Andreas Maier[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen 91058, Germany
siyuan.mei@fau.de
[2] TReNDS Center, Georgia State University, Atlanta 30303, USA
yntn32@outlook.com
[3] Department of Orthodontics and Orofacial Orthopaedics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen 91054, Germany
[4] UTHealth Houston, Houston 77030, TX, USA
[5] Digital Technology and Innovation, Siemens Healthineers, Shanghai 201318, China
[6] Institute of Medical Technology, Peking University, Beijing 100191, China

Abstract. Recent advances in diffusion and flow models have improved the quality and efficiency of image translation. However, their commonly cited advantages, such as diversity sampling and iterative denoising, may not be considered in the medical domain, where fidelity and quantitative accuracy are paramount. Returning to the essence, the diffusion and flow frameworks cast the distribution transport problem as a continuous-time process, injecting the time variable t as a condition into both the network and the input image. In this paper, we rethink the role of this time-conditioned mechanism in medical image translation and remove other diffusion components in training and inference. Our clean setup yields a powerful one-step deterministic generator. In this context, diffusion is conceptually nothing more than a "Just image Regularizer" with t, or what we call JiR. We support this view with task-driven redesigns of the forward process and the prediction space, and introduce a time-consistency loss that aligns the prediction at any time point with the prediction at the end point. Extensive experiments on cross-modality MR-to-PET and multi-contrast MR translation tasks demonstrate that JiR significantly alleviates overfitting and delivers marked improvements over strong baselines, making diffusion more applicable in this field. We hope our findings motivate the community to revisit the foundations of diffusion and flow models and adopt them more objectively. Our code is available at https://github.com/siyuan-mei/JiR.

Keywords: Medical image translation · Flow models · Diffusion models · One-step Generation · Time-conditioned regularization.

1 Introduction

Medical image translation (MIT) has attracted increasing attention in recent years. Many generative modeling approaches from computer vision, especially diffusion models [8,28,13] and flow models [21,20,2], have been widely adopted in medical imaging and have shown impressive results. A large body of recent work [19,18,5,3] focuses on synthesizing a target modality from available source modalities, providing complementary information without additional scans. Typical applications include synthesizing positron emission tomography (PET) from magnetic resonance (MR) to support assisted diagnosis while avoiding the expense and radiation exposure of PET imaging [18,19], and synthesizing contrast-enhanced MR from non-contrast MR to improve lesion conspicuity in settings where contrast administration is undesirable or contraindicated [5,3].

While promising, MIT is ultimately constrained by fidelity and quantitative correctness, often under restricted requirements of reliability and determinism. This appears at odds with the standard diffusion narrative that attributes benefits to diversity sampling and iterative denoising. Recent observations [25,24,9] suggest that diffusion models can struggle to outperform a simple one-step regression model (RM) in terms of accuracy, while multi-step sampling may introduce hallucinations and replicate undesirable acquisition noise. This naturally raises a fundamental question: Do diffusion/flow models offer any real advantage over regression in fidelity-driven MIT?

In this work, we attempt to answer this question from an easily overlooked perspective: the time condition t, which introduces a key factor absent from standard regression baselines. From first principles, diffusion and flow models represent transport between a source distribution π_0 and a target distribution π_1 as a continuous-time process, modeled by stochastic differential equations (SDEs) and ordinary differential equations (ODEs), respectively. Theoretically, one can define an arbitrary time-dependent trajectory $\{\boldsymbol{x}_t\}_{t\in[0,1]}$ to bridge paired $\boldsymbol{x}_0 \sim \pi_0$ and $\boldsymbol{x}_1 \sim \pi_1$, as long as the instantaneous velocity field can be estimated by training a neural network [21,1].

In pursuit of deterministic generation, a one-step ODE sampler is used to directly map the source image to the target, without introducing stochastic noise during iterative sampling [5,25]. To this end, only time-conditioned training is preserved: instead of learning a single mapping $T : \boldsymbol{x}_0 \mapsto \boldsymbol{x}_1$, we train a shared predictor over a continuum of intermediate states, $f_\theta(\boldsymbol{x}_t, t) \mapsto \boldsymbol{x}_1$. Viewed this way, diffusion/flow modeling contributes a simple yet powerful structural bias beyond regression: the time-conditioned formulation regularizes learning by coupling predictions across a continuous-time trajectory. Crucially, this regularization is fundamentally different from noise-based data augmentation, which involves only random input perturbations.

Based on this concept, our approach is nothing more than a "*Just image Regularizer*" (JiR) with t, which distills diffusion-style modeling into a one-step MIT framework. Notably, the proposed JiR is architecture-agnostic and features task-driven redesigns along three axes: (i) the forward interpolant that defines $\{x_t\}$, (ii) the prediction space used for direct regression supervision, and (iii)

a time-consistency loss that anchors the prediction at any time point to the trajectory's end point. Systematic studies on cross-modality MR-to-PET and multi-contrast MR translation show that JiR achieves competitive results while significantly alleviating overfitting compared to strong regression baselines.

2 Related Work

Traditional generative models, particularly conditional generative adversarial nets (GANs) [11,33], have long dominated MIT. To match the volumetric nature of most medical image modalities, numerous GAN variants are extended to 3D space [30], mostly by incorporating adapted architectural components and auxiliary geometry-guided objectives [31,22]. However, GANs are known to suffer from training instability and mode collapse due to the minimax algorithm, which can require laborious hyperparameter tuning and does not transfer reliably across model architectures and datasets.

On the other hand, diffusion models and newly popular flow models emerge as another line of techniques. By leveraging the mathematical formulations of SDEs/ODEs, these models typically surpass GANs in image generation, both in terms of image quality and diversity, and avoid instability issues [6]. The most representative examples are the denoising diffusion probabilistic models (DDPMs) [8], which use a Gaussian prior. Moreover, variants such as diffusion bridge models (DDBMs) [32,15] and flow matching (FM) [20,21] have been explored to replace initial Gaussian noise with empirical observations from the source distributions. Despite their promising claims, more recent evidence [24,25,9] suggests that these models are indeed inferior to simple one-step regression baselines on MIT benchmarks. This is because standard multi-step diffusion sampling (whether SDE or ODE) inevitably replicates stochastic noise during the iterative process, resulting in image distortion. Our work further avoids these pitfalls by proposing task-targeted redesigns.

3 Methods

Our method adopts the FM framework [20,21], an ODE-based diffusion formulation that is widely used as a modern generative paradigm for its good performance and convenient implementation. In this section, we first provide a brief overview of the original formulation and then introduce our redesigns for MIT.

3.1 Background of Flow Matching

Given two different distributions π_0 and π_1, FM seeks to train a continuous-time model capable of progressively transferring data $\boldsymbol{x}_0 \sim \pi_0$ to $\boldsymbol{x}_1 \sim \pi_1$. Conventionally, the source sample $\boldsymbol{x}_0$ can be either a Gaussian noise or a paired image. During training, an intermediate state $\boldsymbol{x}_t$ at time $t \in [0, 1]$ is produced through linear interpolation:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \qquad t \sim \mathcal{U}(0,1). \tag{1}$$

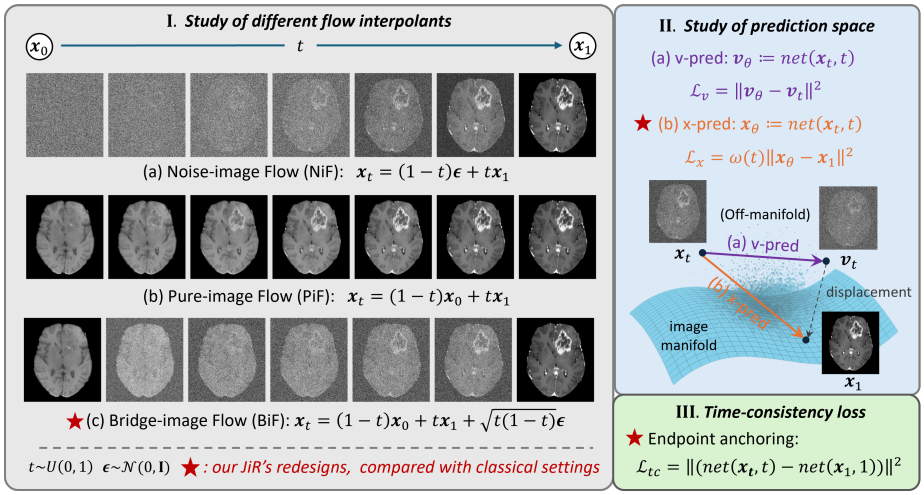

Fig. 1. Our JiR's task-driven redesigns for diffusion training in medical image translation, e.g., multi-contrast MR. We study (I) three flow interpolants for the forward process and (II) the network prediction space based on the manifold hypothesis [16,34], while introducing (III) a time-consistency loss. See text for detailed description.

The instantaneous velocity of the flow $\boldsymbol{v}_t$ is defined as the time-derivative of $\boldsymbol{x}_t$:

$$\boldsymbol{v}_t := \frac{d\boldsymbol{x}_t}{dt} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_t}{1-t} = \boldsymbol{x}_1 - \boldsymbol{x}_0. \tag{2}$$

In practice, $\boldsymbol{v}_t$ can be parameterized by a neural network $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ with a least-square regression loss, which we term as the $\boldsymbol{v}$-*loss*:

$$\mathcal{L}_v = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1} \left\| \boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{v}_t \right\|^2. \tag{3}$$

During inference, samples are obtained by solving the ODE $d\boldsymbol{x}_t/dt = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ from $t = 0$ to $t = 1$ in multiple steps using a numerical solver, usually a first-order Euler or a second-order Heun method. Following a deterministic setup [25,5], we employ one-step Euler to generate $\hat{\boldsymbol{x}}_1 = \boldsymbol{x}_0 + \boldsymbol{v}_\theta(\boldsymbol{x}_0, 0)$.

3.2 Redesigns of Flow Interpolants

The choice of forward interpolants determines how the training data is processed by t and the information received by the network. Following [1], we extend Eq. 1 to a unified stochastic linear interpolant by fixing $\boldsymbol{x}_0$ as the source image and explicitly adding a stochastic term $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$:

$$\boldsymbol{x}_t = \alpha(t)\boldsymbol{x}_0 + \beta(t)\boldsymbol{x}_1 + \gamma(t)\epsilon, \qquad t \sim \mathcal{U}(0,1). \tag{4}$$

As shown in Fig. 1-I, we consider three types of flow interpolants. First, we refer to the two classical flows as **noise-image flow (NiF)** and **pure-image**

flow (PiF), such as NiF sets $\alpha(t) = 0$, $\beta(t) = t$, $\gamma(t) = 1 - t$ and PiF sets $\alpha(t) = 1 - t$, $\beta(t) = t$, $\gamma(t) = 0$. However, for NiF, static source images must be concatenated as a conditional channel, thus lacking the translation dynamic for process modeling; meanwhile, it biases the task towards denoising and introduces stochasticity when sampling starts from random noise. On the other hand, PiF forms a valid transport between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, but produces spurious intermediate modes and sharp density [1].

Inspired by diffusion bridges [15,29], we introduce a simple Brownian bridge into the linear interpolation, which forms a **Bridge-image flow (BiF)**:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1 + \sqrt{t(1-t)}\,\boldsymbol{\epsilon}. \tag{5}$$

Such, $\boldsymbol{v}_t$ in Eq. 2 and Eq. 3 converts to $\boldsymbol{v}_t = (\boldsymbol{x}_1 - \boldsymbol{x}_0) - \sqrt{\frac{t}{1-t}}\epsilon$ [29]. From a statistical learning perspective, this added noise variable effectively smooths both the intermediate mode and the velocity spatially and reaches its maximum at $t = 0.5$. Moreover, it eliminates the formation of spurious features while maintaining the influence of $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ everywhere except at the endpoints [1].

3.3 Redesigns of Prediction Space

Although velocity estimation is commonly used, we find direct network prediction in image space is an easier way of learning, as suggested in [16,26]. According to the manifold hypothesis, adding perturbations will push images away from the low-dimensional space [34]. As visualized in Fig. 1-II, since the velocity field $\boldsymbol{v}_t$ essentially is a displacement with noise, the $\boldsymbol{v}$-prediction rests on the off-manifold space, which obviously increases the learning difficulty. Therefore, we let the network directly predict the target image $\boldsymbol{x}_1$, which we term $\boldsymbol{x}$-prediction [16,26] with the *$\boldsymbol{x}$-loss*:

$$\mathcal{L}_x = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\omega(t)\left\|\boldsymbol{x}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_1\right\|^2, \tag{6}$$

where $\boldsymbol{x}_\theta := net(\cdot,\cdot)$ and $\omega(t) = \frac{1}{4t(1-t)}$ is a t-dependent scale (we clip its denominator to 0.05 to prevent zero division). Notably, compared to the $\boldsymbol{x}$-loss in unconditional image generation where $\omega(t) = \frac{1}{(1-t)^2}$ [16], we increase the weights when t is close to 0 and reduce the oscillation amplitude, making training more stable and focused on translation and reconstruction rather than denoising.

3.4 Time-Consistency Loss

To further enhance determinism, we leverage a self-consistency property in the ODE trajectory [28,27], *i.e.*, the prediction at any time point should align with the prediction at the end point. Specifically, we set the end point $t = 1$ as an anchor and introduce a time-consistency loss (*tc-loss*):

$$\mathcal{L}_{tc} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\left\|\boldsymbol{x}_\theta(\boldsymbol{x}_t, t) - \text{stopgrad}(\boldsymbol{x}_\theta(\boldsymbol{x}_1, 1))\right\|^2, \tag{7}$$

where the "stopgrad" operator means no gradient backpropagated from the anchor point. This objective is similar in concept to the Consistency Model

(CM) [27], but the trajectory of CM is anchored by adjacent points. In our case, choosing the end point is more reasonable since the reconstruction $\boldsymbol{x}_\theta(\boldsymbol{x}_1, 1) \approx \boldsymbol{x}_1$ readily learns an identity mapping. Therefore, chaining predictions across the flow trajectory prevents the model from exploiting t as a shortcut to memorize training samples, instead encouraging smooth and consistent mappings. The final training objective of JiR combines two terms:

$$\mathcal{L}_{\text{JiR}} = \mathcal{L}_x + \lambda_{tc}\mathcal{L}_{tc}, \tag{8}$$

where λ_{tc} is a hyperparameter balancing the two losses (default $\lambda_{tc} = 0.5$). As we will demonstrate, our redesigned system is simple yet effective, significantly outperforming multi-step diffusion baselines.

4 Experiments

We conduct experiments on MR-to-PET translation (Task 1) and multi-contrast MR translation (Task 2). We first present a systematic study and ablation in Section 4.2, and then compare with existing SoTA methods in Section 4.3.

4.1 Datasets and Implementation

MR-to-PET on ADNI We use 829 paired T1-weighted (T1w) MR and ^{18}F-fluorodeoxyglucose (FDG) PET scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[7] [12]. Four diagnostic labels are included: cognitively normal (CN, $n = 239$), Alzheimer's disease (AD, $n = 212$), progressive mild cognitive impairment (pMCI, $n = 129$), and stable mild cognitive impairment (sMCI, $n = 249$). Both modalities are spatially cropped to $128 \times 128 \times 128$.
Multi-contrast MR on BraTS The Brain Tumor Segmentation (BraTS) 2023 dataset [14] contains 1251 multi-contrast MR volumes acquired from glioma patients. We select T1-weighted (T1w) as the source domain and contrast-enhanced T1w (T1ce) as the target domain. Volumes are center-cropped to $128 \times 192 \times 192$ and subsequently resampled to $96 \times 160 \times 160$ for memory efficiency.

Both datasets are co-registered and split into $7 : 1 : 2$ ratios for training, validation, and testing. For preprocessing, we apply min-max normalization and rescale intensities to $[-1, 1]$.
Implementation Since our framework is architecture-agnostic, we adopt the 3D ResUNet [10] as the standard backbone, with time-embeddings injected into each block using feature-wise linear modulation [23]. Training uses the Adam optimizer with a constant learning rate of $2e-4$. All experiments are conducted on an NVIDIA H100 GPU with a batch size of 4. Moreover, we deliberately train the models for 500 epochs to observe their regularization performance and save the checkpoint with the lowest MSE loss in the validation set for testing.

[7] adni.loni.usc.edu

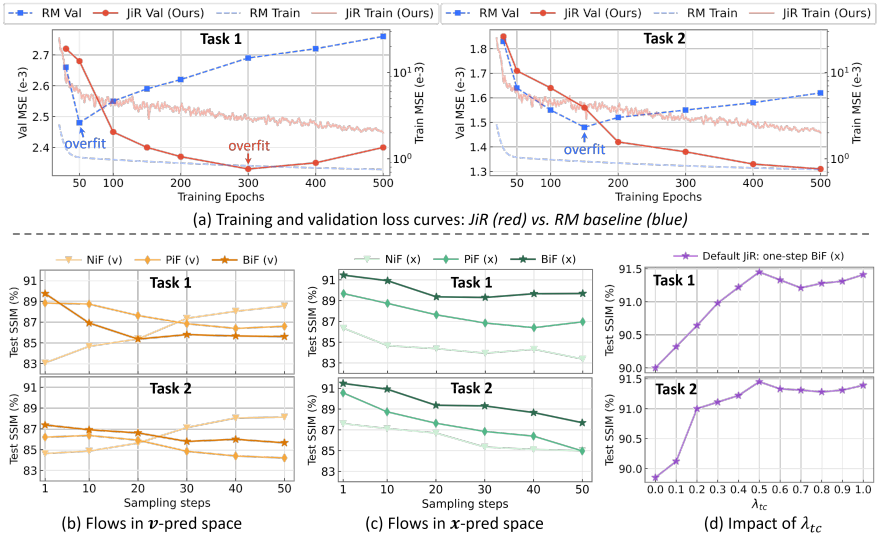

Fig. 2. (a) JiR significantly alleviates overfitting compared to the regression model (RM) across two tasks. (b-d) Ablation study of flow types, prediction space, and loss weight λ_{tc}. Here, we save the best checkpoints on the validation set and measure SSIM on the test set. Due to the page limit, comprehensive metrics are presented in Table 1.

4.2 Systematic Study and Ablation

Fig. 2(a) demonstrates the effect of regularization on training behavior by comparing JiR with its regression baseline (RM) under the same architecture. The only difference is that JiR is trained with time conditions t. It can be seen that the validation loss of RM (blue curves) stops decreasing at epoch 50 for task 1 and at epoch 150 for task 2, reflecting rapid overfitting under typical medical data sizes. Instead, our JiR not only achieves lower validation loss (red curves), but also continues to decrease it over a much longer period: until 300 epochs for task 1 and 500 epochs for task 2. This suggests that time-conditioning acts as an implicit regularizer: an image pair is no longer used to supervise only the endpoint mapping, but a continuum of intermediate conditions, which effectively enhances generalization and discourages shortcut memorization.

Next, Fig. 2(b-d) ablates our three core components described in Section 3. The subfigures (b) and (c) compare three flow interpolants in $\boldsymbol{v}$-prediction and $\boldsymbol{x}$-prediction with 0-50 sampling steps, respectively. For $\boldsymbol{v}$-prediction in Fig. 2(b), only classical NiF improves test SSIM by increasing the sampling steps. Surprisingly, $\boldsymbol{x}$-pred largely surpasses its $\boldsymbol{v}$-counterpart across three flows with only one sampling step, highlighting the importance of task-aligned prediction space. Our BiF(x) reaches the overall highest SSIM of 91.45% for task 1 and 91.46% for task 2. In addition, Fig. 2(d) shows the impact of hyperparameter λ_{tc} in Eq. 8. Adding

Table 1. Quantitative results. "NFE" measures the number of function evaluations. For FM [20] and DDBM [32], we use a standard 50-step Heun sampler by default [13].

Methods	NFE↓	Task 1: MR-to-PET				Task 2: T1w-to-T1ce MR			
		MAE↓	PSNR↑	SSIM↑	MS-SSIM↑	MAE↓	PSNR↑	SSIM↑	MS-SSIM↑
Pix2Pix3D [11]	1	0.0239 ±0.0055	24.98 ±1.56	87.92 ±1.82	91.70 ±1.58	0.0173 ±0.0078	27.90 ±2.18	87.21 ±2.88	87.41 ±3.21
Ea-GAN [31]	1	0.0219 ±0.0058	25.80 ±1.85	89.25 ±2.20	92.87 ±1.75	0.0153 ±0.0079	29.46 ±2.90	90.55 ±3.16	<u>91.72</u> ±3.33
SwinUNETR [7]	1	0.0221 ±0.0057	25.84 ±1.84	89.95 ±1.93	93.05 ±1.59	0.0158 ±0.0086	29.07 ±2.70	90.49 ±3.26	90.97 ±3.27
ResUNet [10] (RM baseline)	1	<u>0.0214</u> ±0.0056	<u>26.12</u> ±1.89	<u>90.11</u> ±1.96	<u>93.32</u> ±1.61	<u>0.0152</u> ±0.0088	<u>29.49</u> ±2.96	<u>90.74</u> ±3.72	91.59 ±3.33
FM [20]	99	0.0228 ±0.0061	25.65 ±1.90	88.55 ±2.14	92.75 ±1.67	0.0170 ±0.0086	28.43 ±2.63	88.13 ±3.20	89.38 ±3.34
DDBM [32]	99	0.0281 ±0.0051	25.33 ±1.55	75.68 ±4.19	92.18 ±1.50	0.0281 ±0.0082	24.62 ±1.84	74.43 ±6.69	83.14 ±4.17
JiR (Ours)	1	**0.0205** ±0.0055	**26.61** ±1.95	**91.45** ±2.13	**94.36** ±1.66	**0.0145** ±0.0077	**30.15** ±2.98	**91.46** ±2.37	**92.40** ±2.86

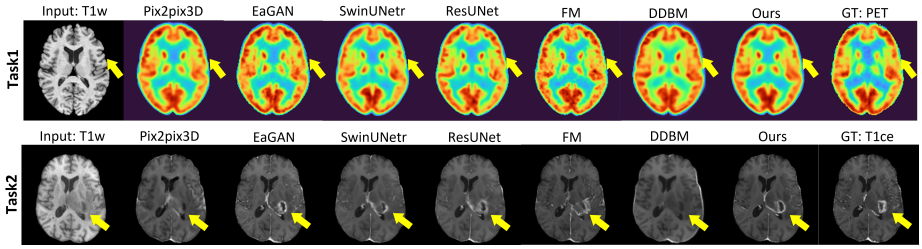

Fig. 3. Qualitative results of comparison study on two tasks. Yellow arrows highlight regions where the methods differ.

the time-consistency loss consistently improves performance, and $\lambda_{tc} = 0.5$ provides a sweet spot that strengthens accuracy without over-constraining.

4.3 Comparison with SoTA

Table 1 compares JiR with GAN-based translators (Pix2Pix3D [11], Ea-GAN [31]), strong regression models (SwinUNETR [7], ResUNet [10]), and recent diffusion-based methods (FM [20], DDBM [32]). JiR achieves the best performance across all four fidelity metrics on both tasks while keeping one-step inference (NFE=1); notably, none of the other competing paradigms outperforms the plain ResUNet regression baseline (RM). On MR-to-PET, JiR improves over RM from MAE $0.0214 \rightarrow 0.0205$ and SSIM $90.11\% \rightarrow 91.45\%$ (MS-SSIM $93.32\% \rightarrow 94.36\%$). On T1w-to-T1ce, JiR further reduces MAE $0.0152 \rightarrow 0.0145$ and increases PSNR $29.49 \rightarrow 30.15$ and SSIM $90.74\% \rightarrow 91.46\%$. Moreover, multi-step FM/DDBM with a 50-step Heun solver (NFE=99) does not yield better accuracy, and DDBM exhibits severe SSIM degradation. Fig. 3 qualitatively confirms that JiR better preserves structures and avoids over-smoothing/artifacts in highlighted regions.

5 Conclusion

Our work revisits the functional requirements of diffusion models in the problem of MIT. Through fair, architecture-matched comparisons against strong regression baselines, we show that the time condition t is the key differentiator of diffusion-style training, and that this mechanism serves as an effective form of regularization. Building on this understanding, we redesign the entire system from multiple perspectives to improve the quality of one-step translation, substantially reducing computational overhead without sacrificing fidelity. Interestingly, the proposed JiR echoes recent research on the generalization and regularization of diffusion models in other fields [4,17], offering an exciting cross-pollination for the medical imaging techniques.

Acknowledgments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

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

References

1. Albergo, M., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. Journal of Machine Learning Research **26**(209), 1–80 (2025)
2. Albergo, M., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: ICLR 2023 Conference (2023)
3. Arslan, F., Kabas, B., Dalmaz, O., Ozbey, M., Çukur, T.: Self-consistent recursive diffusion bridge for medical image translation. Medical Image Analysis **106**, 103747 (2025)
4. Bonnaire, T., Urfin, R., Biroli, G., Mézard, M.: Why diffusion models don't memorize: The role of implicit dynamical regularization in training. Advances in Neural Information Processing Systems **38**, 141266–141286 (2026)
5. Chang, H., Shang, Y., Wang, H., Liang, Y., Wang, H., Wang, F., Niu, C., Lian, C.: Controllable flow matching for 3d contrast-enhanced brain mri synthesis from non-contrast scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 119–128. Springer (2025)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Huijben, E.M., Terpstra, M.L., Pai, S., Thummerer, A., Koopmans, P., Afonso, M., Van Eijnatten, M., Gurney-Champion, O., Chen, Z., Zhang, Y., et al.: Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report. Medical image analysis **97**, 103276 (2024)

10. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
11. 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)
12. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., 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)
13. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)
14. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation (brats) challenge 2023: focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). ArXiv pp. arXiv–2305 (2024)
15. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023)
16. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
17. Li, X., Zhang, Z., Li, X., Chen, S., Zhu, Z., Wang, P., Qu, Q.: Understanding representation dynamics of diffusion models via low-dimensional modeling. Advances in Neural Information Processing Systems **38**, 107365–107404 (2026)
18. Li, Y., Buchert, R., Schmitz-Koep, B., Grimmer, T., Ommer, B., Hedderich, D.M., Yakushev, I., Wachinger, C.: Diffusion bridge networks simulate clinical-grade pet from mri for dementia diagnostics. arXiv preprint arXiv:2510.15556 (2025)
19. Li, Y., Yakushev, I., Hedderich, D.M., Wachinger, C.: Pasta: Pathology-aware mri to pet cross-modal translation with diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–540. Springer (2024)
20. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
21. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
22. Mei, S., Xia, Y., Fan, F., Maier, A.: Ganext: A fully convnext-enhanced generative adversarial network for mri-and cbct-to-ct synthesis. arXiv preprint arXiv:2512.19336 (2025)
23. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
24. Pfaff, L., Wagner, F., Vysotskaya, N., Thies, M., Maul, N., Mei, S., Wuerfl, T., Maier, A.: No-new-denoiser: A critical analysis of diffusion models for medical image denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 568–578. Springer (2024)

25. Rassmann, S., Kügler, D., Ewert, C., Reuter, M.: Regression is all you need for medical image translation. IEEE Transactions on Medical Imaging (2026)
26. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
27. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
28. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
29. Tan, Z., Wang, Z., Yang, X., Liu, S., Wang, X.: Vision bridge transformer at scale. arXiv preprint arXiv:2511.23199 (2025)
30. Usama, Z., Alavi, A., Chan, J.: A review on the applications of gans for 3d medical image analysis. Applied Sciences **15**(20), 11219 (2025)
31. Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P.: Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis. IEEE transactions on medical imaging **38**(7), 1750–1762 (2019)
32. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: International Conference on Learning Representations. vol. 2024, pp. 8160–8171 (2024)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
34. Zhu, X.: Semi-supervised learning literature survey. world **10**, 10 (2005)