Posterior Samplings are Missing Modalities Generators for Medical Image Translation

Jonghun Kim

Department of Electrical and Computer Engineering,
Sungkyunkwan University, Suwon, Korea
iproj2@skku.edu

Abstract. Magnetic resonance imaging comes in various modality contrasts that provide complementary anatomical and pathological information. Complete multimodal acquisitions are often unavailable due to time and protocol constraints. This leads to real-world datasets with missing modalities, where conventional medical image translation methods are typically limited to fixed source-target settings or require retraining for each observed source-target pair. We propose a unified framework that formulates missing-modality generation as a linear inverse problem under a joint distribution and solves it via posterior sampling with a flow matching model. By learning a joint prior over the complete modality set, our method can reconstruct arbitrary missing modalities at inference time by guiding the sampling trajectory to enforce measurement consistency with observed modalities. We further mitigate inter-modality error propagation in multi-target generation by adopting a many-to-one sampling strategy. Experiments on BraTS and IXI datasets show that our method achieves the best performance over baselines across most missing-modality scenarios. In downstream tumor segmentation, synthesized images from our method result in higher segmentation performance, indicating better preservation of clinically relevant structures. Our code is available at github.com/jongdory/PS-MIT.

Keywords: Medical Image Translation · Posterior Sampling · Flow Matching

1 Introduction

Magnetic resonance imaging (MRI) is an essential non-invasive tool in clinical workflows [25]. MRI comes in multiple modalities, such as T1-weighted, T2-weighted, and T2-weighted fluid-attenuated inversion recovery (FLAIR) images, that provide complementary anatomical and pathological information [35]. For example, T1-weighted images emphasize anatomy and tissue boundaries, while T2-weighted and FLAIR images provide contrast for identifying pathological changes, such as edema or lesions, by highlighting differences in water content [13, 16]. Analyzing these multimodal images is necessary for a comprehensive understanding of a patient's condition. Despite their clinical utility, acquiring a complete set of modalities is impractical due to time and protocol constraints.

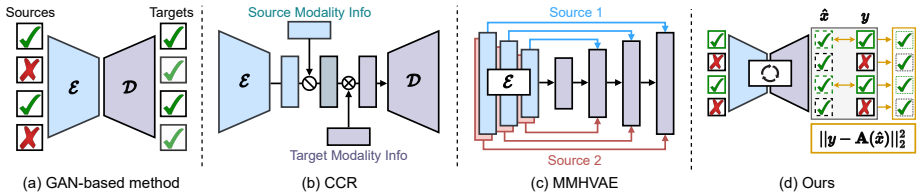

Fig. 1: Overview of representative missing-modality generation frameworks. **(a) GAN-based method** [43] concatenates observed modalities channel-wise and handles missing channels with zeros. **(b) CCR** [46] removes source information from encoder features and injects target information. **(c) MMHVAE** [14] generates targets from latent variables using features from modality-specific encoders. **(d) Ours** uses a flow-based model and posterior sampling. The sampling trajectory is guided by the current estimate $\hat{\boldsymbol{x}}$ and measurements $\boldsymbol{y}$ to enforce consistency with observed modalities.

For instance, FLAIR requires long repetition and inversion times to suppress cerebrospinal fluid signals [30, 35]. As a result, real-world datasets often contain incomplete data where certain modalities are missing.

To address the missing-modality problem, medical image translation has been studied with generative models, evolving from generative adversarial network (GAN)-based methods to diffusion-based methods [2, 11, 27, 29, 36, 47]. However, existing work focuses on fixed translation settings, translating a commonly available source modality into a frequently missing target modality. Such fixed formulations do not reflect the diverse missing-modality scenarios encountered in practice and often require training separate models for different input-output pairs. In fixed settings, if a specific required modality is missing, models cannot perform synthesis and cannot benefit from additional available modalities that potentially improve generation quality. Consequently, these methods are inflexible and inefficient to deploy across tasks. An ideal framework should support synthesis from arbitrary observed modalities within a unified model and dynamically fuse information across available modalities [14, 43, 46].

To address these requirements, unified approaches such as MM-GAN [43], ResViT [11], CCR [46], and MMHVAE [14] have been proposed to generate missing modalities from partially observed inputs (Fig 1). For instance, MM-GAN [43] and ResViT [11] rely on zero imputation to handle incomplete data. CCR [46] attempts to explicitly disentangle contrast and content representations, but in medical imaging, these two components are highly related. MRI contrast depends on the intrinsic relaxation times of specific anatomical structures. Thus, decoupling them could result in the loss of critical pathological features, such as tumor edema [6]. Forcing this separation risks losing inherent correlations causing structural distortions in the synthesized images. MMHVAE [14] utilizes a multi-level hierarchical structure to generate high-resolution images, which requires a complex design to compress and reconstruct information at each layer managing dependencies between latent variables. Furthermore, the optimization of multiple sub-networks and the hierarchical structure makes it challenging to balance the various loss terms.

Unlike previous methods that depend on complex architectural or explicit disentanglement, we treat missing modality generation as a linear inverse problem under a joint distribution and solve it via posterior sampling [7, 45] with an iterative generative model [22, 31, 32, 44]. We train a flow-based model [31] to learn the joint prior over the complete modality set. At inference time, we start from noise and guide the sampling trajectory using the observed modalities to reconstruct missing modalities. Our method has the advantage of addressing all missing scenarios by learning a standard joint prior. By learning a single joint prior, our method supports diverse observed-missing combinations and leverages all available modalities.

Our contributions are as follows:

- We propose a novel approach for missing modality generation by training a joint generative model and employing posterior sampling.
- We identify the error accumulation when generating multiple modalities simultaneously and propose a many-to-one sampling strategy to mitigate this.
- Our approach successfully generates missing modalities and outperforms baseline models on the BraTS and IXI datasets.

2 Related Works

Medical Image Translation. Medical image translation aims to synthesize missing modalities. Prior work has adopted GANs [18] and, more recently, diffusion models [22]. Early GAN-based approaches primarily relied on the Pix2Pix [24], with notable advancements such as Ea-GAN [47] for edge guidance, ResViT [11] for transformer-based architectures, and RegGAN [29] for addressing spatial misalignments. More recently, diffusion-based models have emerged, including ALDM [27], which utilizes style information within the latent space. SynDiff [36] integrates the diffusion process with adversarial training. SelfRDB [2] employs a diffusion-bridge mechanism. However, these models are typically constrained to fixed one-to-one or many-to-one settings. To support diverse missing-modality scenarios (Fig. 1), unified frameworks such as MM-GAN [43], CCR [46], and MMHVAE [14] have been proposed. They broaden the translation setting beyond fixed pairs but often introduce additional architectural or optimization complexity. In contrast, our approach avoids specialized architectures by learning a joint generative prior over the complete multimodal set and reconstructing missing modalities via posterior sampling at inference time.

Inverse Problems. Inverse problems aim to recover a true signal $\boldsymbol{x}$ from a measurement $\boldsymbol{y}$ corrupted by a forward model. This forward model usually represents a physical, computational, or statistical process that maps the true signal to the measurement space. This process often introduces noise, distortion, or information loss. Recent methods leverage pre-trained diffusion and flow-based models as generative priors $p(\boldsymbol{x})$. By integrating the forward model and measurement, these approaches sample from the posterior $p(\boldsymbol{x}|\boldsymbol{y})$, often utilizing the log-likelihood gradient $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x})$, to solve tasks like inpainting and super-resolution without task-specific retraining [12]. Early works focused on diffusion

models that operate in pixel space [7–9, 45]. These methods were later extended to latent space to take advantage of pre-trained latent models [10, 42, 49]. More recently, flow-based models have been proposed for solving linear inverse problems [26, 38]. DPS [7] guides the reverse diffusion process using an approximation of the likelihood gradient to improve measurement consistency. DDNM [45] enforces data consistency via a projection-based decomposition into range and null spaces. PSLD [42] applies gradient-based guidance in the latent space to enable efficient high-resolution restoration. These posterior-sampling methods have also been extended to flow models. FlowChef [38] introduces a guidance term into the reverse ordinary differential equation (ODE) to enforce data consistency, while FlowDPS [26] generalizes posterior sampling to affine flows by decomposing Euler steps using a generalized Tweedie's formula [15]. These methods can be categorized by their constraint mechanisms. DPS, FlowChef, and FlowDPS use gradient-based soft constraints. In contrast, DDNM uses a projection-based hard constraint to enforce measurement consistency. PSLD uses a hybrid approach that combines gradient-based and projection-based constraints. While previous studies have primarily focused on improving the perceptual quality in natural image domain, our goal is to improve image fidelity in the medical domain. We compare the performance of various posterior sampling methods to identify the optimal method.

3 Preliminaries

3.1 Conditional Flow Matching.

Conditional flow matching (CFM) learns a probability path p_t that transports a noise distribution p_0 to a data distribution q via a straight-line trajectory. We define $\boldsymbol{x}_t$ as:

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

where $t \in [0,1]$, $\boldsymbol{x}_0 \sim p_0$ is noise and $\boldsymbol{x}_1 \sim q$ is a target sample. The corresponding conditional vector field is $u_t(\boldsymbol{x}_1|\boldsymbol{x}_0) = \boldsymbol{x}_1 - \boldsymbol{x}_0$. Our model takes a multi-channel image $\boldsymbol{x}_t$, where each channel corresponds to a different modality. The training objective is:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(\boldsymbol{x}_1),p_0(\boldsymbol{x}_0)} \|u_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2, \tag{2}$$

where θ denotes the learnable parameters of the neural network for u_θ and t is sampled uniformly from $\mathcal{U}(0,1)$. This objective trains the vector field that transforms a Gaussian distribution into a multimodal data distribution.

3.2 Problem Formulation

The task of reconstructing missing modalities can be mathematically formulated as a linear inverse problem. A typical linear inverse problem aims to recover an unknown clean sample $\boldsymbol{x}_1 \in \mathbb{R}^d$ from a set of measurements $\boldsymbol{y}$:

$$\boldsymbol{y} = \mathcal{A}\boldsymbol{x}_1 + \boldsymbol{n} \in \mathbb{R}^m, \tag{3}$$

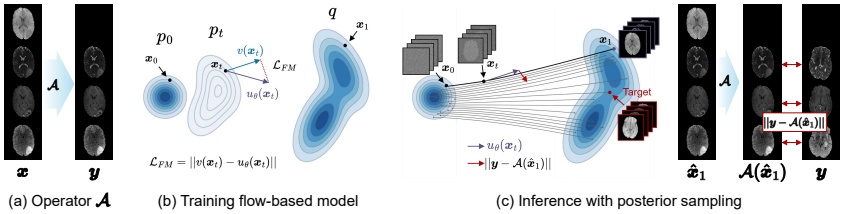

Fig. 2: Overview of the proposed training and inference pipeline. **(a)** The measurement operator $\mathcal{A}$ masks missing modality channels and $\boldsymbol{y}$ denotes the observed modalities. **(b)** The flow-based model learns a time-dependent vector field that transports noise to data. **(c)** At inference, missing modalities are reconstructed via posterior sampling guided to minimize the difference between the projected estimate $\mathcal{A}(\hat{\boldsymbol{x}}_1)$ and $\boldsymbol{y}$.

where $\mathcal{A} \in \mathbb{R}^{m \times d}$ is the measurement operator and $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ denotes Gaussian noise. Inverse problems are inherently ill-posed. There are multiple solutions $\boldsymbol{x}$ that satisfy Eq. 3. We represent a multimodal sample as a multi-channel image $\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$, where C denotes the number of modalities. The masking operator $\mathcal{A}$ selects the observed input modality channels. In our work, $\mathcal{A}$ is a channel-selection operator determined by the available modalities at inference time and $\boldsymbol{y}$ contains only observed channels.

4 Methods

We formulate missing-modality generation as a linear inverse problem and solve it using a pre-trained flow-based model. Specifically, we learn a joint generative model over multi-modal images by representing all modalities as channels of a single multi-channel sample. At inference time, missing modalities are reconstructed via posterior sampling (Fig. 2). The model learns a probability path from noise to data by predicting the corresponding vector field at each time step t. During sampling, we enforce measurement consistency using the residual between the projected estimate $\mathcal{A}(\hat{\boldsymbol{x}}_1)$ and the measurement $\boldsymbol{y}$, guiding the trajectory to synthesize missing channels while preserving observed modalities.

4.1 Measurement-Consistent Sampling

In medical image translation, fidelity is the main priority and takes precedence over perceptual quality or diversity [39]. Any structural inconsistency or hallucination in the synthesized output can severely compromise its clinical utility. To mitigate these risks and ensure anatomical accuracy, we employ a sampling strategy that enforces strict data consistency. Flow-based models generate samples by solving the probability flow ODE. The sampling update is defined as:

$$d\boldsymbol{x}_t = v_\theta(\boldsymbol{x}_t, t)\, dt, \qquad \boldsymbol{x}_{t+dt} = \boldsymbol{x}_t + v_\theta(\boldsymbol{x}_t, t)\, dt, \tag{4}$$

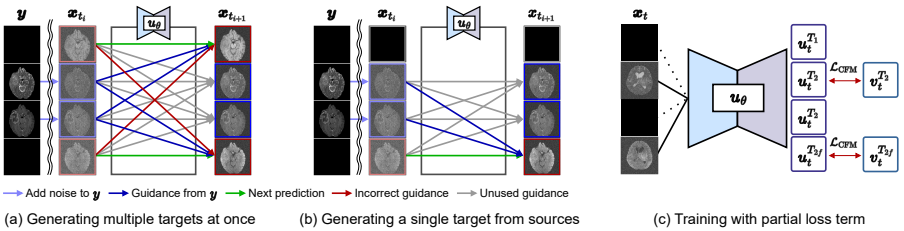

Fig. 3: Mitigating inter-modality error propagation. **(a)** Jointly synthesizing multiple modalities may propagate errors across channels (red arrows). **(b)** Many-to-one sampling generates one target modality at a time, avoiding interference from other synthesized channels. **(c)** For training, our model uses random combinations of observed-target scenarios and calculates the vector field of the observed channels.

where v_θ denotes the vector field predicted by the model. To reconstruct missing modalities while preserving the observed modalities, we introduce two interventions during sampling. First, we anchor the observed channels by replacing the predicted vector field in those channels with a measurement-induced vector field. Following inpainting-style strategies, we enforce the ideal velocity toward the ground truth in the observed channels, while letting the unobserved (missing) channels evolve according to the model prediction [33, 34]. Let C be the number of modality channels and let $\boldsymbol{y}$ denote the observed measurements. We use a binary channel mask $\boldsymbol{m} \in \{0,1\}^C$ to indicate which channels are observed or missing. The modified vector field $\hat{v}_t$ used for the update step is defined as:

$$\hat{v}_t = \boldsymbol{m} \odot (\boldsymbol{y} - \boldsymbol{x}_0) + (1 - \boldsymbol{m}) \odot u_\theta(\boldsymbol{x}_t, t). \tag{5}$$

This keeps observed channels consistent, while allowing the model to synthesize missing channels. Second, we apply a posterior-sampling guidance term minimizing a $\|\boldsymbol{y} - \mathcal{A}(\hat{\boldsymbol{x}}_1)\|_2^2$, which steers the trajectory toward the measurement-consistent set:

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - s\nabla_{\hat{\boldsymbol{x}}_1}\|\boldsymbol{y} - \mathcal{A}(\hat{\boldsymbol{x}}_1)\|_2^2, \tag{6}$$

where $\nabla_{\hat{\boldsymbol{x}}_1}$ is the gradient of $\hat{\boldsymbol{x}}_1$ and s denotes the step size.

4.2 Mitigating Inter-Modality Error Propagation

The proposed method can simultaneously synthesize all modalities regardless of the number of missing modalities. However, in scenarios with multiple missing modalities, particularly in many-to-many generation tasks, error accumulation may occur. During generation, deviations from the target trajectory in a given modality can propagate to other modalities, potentially affecting the generation paths of other target modalities. The generation process should rely solely on real, observed information. However, the joint generation process allows irrelevant information or errors originating from an intermediate stage of unverified

modalities to propagate to other target modalities (Fig. 3(a)). These accumulated errors can lead to severe structural inconsistencies. In contrast, as shown in Fig. 3(b), the single-target generation process fully utilizes the observed modality information without incorrect guidance.

Training Strategy. The training process must be flexible to handle diverse target generation scenarios. The model should use only the observed modalities as inputs for each target translation task. For example, in T2→FLAIR scenario (Fig. 3(c)), only the T2 and FLAIR channels remain active while other channels are masked. During the sampling process, FLAIR is generated with the observed modality T2. To handle all scenarios, the model is trained on the vector field of the input modalities. Missing modalities are zeroed to be excluded from the computation, as zero-filling an input channel is equivalent to its removal [17,48]. The convolution operation relies on the multiplication of inputs and kernels, thus setting an input channel to zero is equivalent to pruning its corresponding weight channel. We scale the output based on the number of inputs after the convolution step to prevent variability due to varying numbers of channels. During training, we employ a stochastic masking strategy where up to two input channels are randomly masked. We define a scenario mask $\boldsymbol{m}_s \in \{0,1\}^C$ to indicate which channels are used. The model is optimized using a partial loss as shown in Fig. 3.

Sampling Strategy. For many-to-one sampling, we ensure that only the observed and target modalities influence the generation process with scenario mask $\boldsymbol{m}_s$. For example, if there are two observed modalities and one target modality, all other channels are zeroed so that only these three channels are active. We perform posterior sampling guided by the observed modalities and overwrite the observed vector fields using the mask $\boldsymbol{m}$. The many-to-one sampling procedure is summarized in Algorithm 1. Notably, in the absence of $\boldsymbol{m}_s$, the procedure becomes many-to-many sampling.

Algorithm 1 Many-to-one sampling

Input: Flow-based model u_θ, input noise sample $\boldsymbol{x}_0 \sim N(0, I)$, mask $\boldsymbol{m}$, scenario mask $\boldsymbol{m}_s$, measurements $\boldsymbol{y}$, and step size s.
for $t \in \{0...1\}$ **do**
  $\boldsymbol{x}_t \leftarrow \boldsymbol{m}_s \odot \boldsymbol{x}_t$ // scenario masking
  $v \leftarrow u_\theta(\boldsymbol{x}_t, t)$
  $dt \leftarrow 1/T$
  $\hat{\boldsymbol{x}}_1 \leftarrow \boldsymbol{x}_t + (1-t) \cdot v$
  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_t - s\nabla_{\hat{\boldsymbol{x}}_1} \|\boldsymbol{y} - \mathcal{A}(\hat{\boldsymbol{x}}_1)\|_2^2$ // Eq. 6
  $v^{obs} \leftarrow \boldsymbol{y} - \boldsymbol{x}_0$ // observed vector field
  $v \leftarrow \boldsymbol{m} \odot v^{obs} + (1 - \boldsymbol{m}) \odot v$ // Eq. 5
  $\hat{\boldsymbol{x}}_{t+1} \leftarrow \boldsymbol{x}_t + dt \cdot v$
return $\boldsymbol{x}_1$

Expectation Approximation Sampling. In physical MRI acquisition, noise is inevitable [19]. Generative models learn this inherent noise during training. In practice, repeatedly acquiring MRIs and averaging the signals can effectively suppress this noise [1,40]. Ideally, as the number of acquired samples approaches infinity, the average converges to a noise-free image. Similarly, since flow-based models sample non-deterministically from random noise, each generated sample contains different random noise. Leveraging this, previous research [39] proposed Expectation-Approximation (ExpA) sampling, which adapts the physical principle of signal averaging to generative models to suppress noise. The noise-free image $\boldsymbol{x}'$ can be approximated by averaging multiple generated samples with the

number of ExpA sampling N_{Ex}:

$$\mathbb{E}_{\hat{\boldsymbol{x}}\sim p_\theta} = \lim_{N_{\text{Ex}}\to\infty} \hat{\bar{\boldsymbol{x}}}_{N_{\text{Ex}}} \approx \lim_{N_{\text{Ex}}\to\infty} \bar{\boldsymbol{x}}_{N_{\text{Ex}}} = \boldsymbol{x}', \tag{7}$$

where p_θ is the probability distribution based on networks, $\hat{\boldsymbol{x}}$ is a generated image, $\hat{\bar{\boldsymbol{x}}}$ is the average of generated images, and $\bar{\boldsymbol{x}}$ is the average of measured images.

5 Experiments

5.1 Experimental Setup

Datasets. We used the BraTS2023 [3,4,35] and IXI [1] datasets for model training and validation. From the BraTS, we used T1w, T2w, T1ce, and FLAIR images. A total of 1140 subjects were randomly selected, with 1000 used for training, 40 for validation, and 100 for testing. All images were skull-stripped and resampled to 1 mm isotropic resolution. From the IXI dataset, we used T1w, T2w, and proton density (PD) images. We randomly selected 200 subjects, allocating 120 for training, 20 for validation, and 60 for testing. All images were preprocessed for skull stripping, affine registration, and resampling to a spacing of $0.9375 \times 0.9375 \times 1.2\text{mm}^3$.

Implementation Details. All experiments were performed using PyTorch [37] in Python with CUDA 12.1 and the Adam [28] optimizer was employed for training. A single NVIDIA A100 80GB GPU was used. For the sampling, the number of function evaluations (NFE) was set to 40, while N_{Ex} was set to 10. We adopted PSLD [42] as the posterior sampling method.

Model Architectures. The backbone network architecture was a UNet [41] with both the encoder and decoder comprising four ResNet [20] blocks using 128, 256, 512, and 512 channels.

Metrics. To evaluate image fidelity, PSNR and SSIM were utilized. The perceptual quality was assessed using FID [21] and LPIPS [50] to quantify the perception-distortion trade-off [5].

Comparison Methods. We compared the unified models, MM-GAN [43], ResViT [11], CCR [46], and MMHVAE [14] as baseline models for generating missing modalities. Additionally, we compared various posterior sampling methods of DPS [7], DDNM [45], PSLD [42], FlowChef [38], and FlowDPS [26]. The supplement contains details of the posterior sampling methods.

5.2 Results with Comparison Methods

Results on the BraTS. We evaluated our method across all possible missing modality scenarios on the BraTS, with quantitative results summarized in Table 1. In most cases, the proposed method outperformed the baselines. Performance

[1] https://brain-development.org/ixi-dataset/

Table 1: Quantitative performance of MRI synthesis with baselines on the BraTS. T_1, T_2, T_{1ce}, and T_{2f} denote T1-weighted, T2-weighted, T1 contrast-enhanced, and FLAIR, respectively. Sources denote the observed modalities. Mean and standard deviation values are presented. Blue indicates the best performance. Underline indicates statistically significant differences compared to all others ($p < 0.05$).

Target	Sources				PSNR (dB) ↑					SSIM (%) ↑				
	T_1	T_2	T_{1c}	T_{2f}	MMGAN	ResViT	CCR	MMHVAE	**Ours**	MMGAN	ResViT	CCR	MMHVAE	**Ours**
T_1	○	●	○	○	24.36 (2.63)	23.05 (2.12)	24.02 (2.51)	22.67 (2.49)	25.56 (2.39)	90.28 (3.24)	89.61 (3.37)	89.92 (3.20)	88.10 (3.47)	90.52 (3.12)
	○	○	●	○	25.24 (3.09)	23.86 (2.83)	25.15 (2.63)	23.92 (2.60)	25.56 (2.14)	87.70 (3.27)	87.40 (3.31)	87.59 (3.01)	89.15 (3.57)	88.17 (2.84)
	○	○	○	●	23.96 (2.24)	22.95 (1.60)	23.63 (2.37)	21.49 (2.51)	24.66 (2.61)	88.02 (2.85)	88.01 (2.83)	87.74 (3.02)	84.28 (3.09)	88.27 (3.06)
	○	●	●	○	26.01 (3.26)	24.01 (3.14)	25.50 (2.83)	24.61 (2.85)	27.15 (2.43)	91.94 (3.20)	90.82 (3.52)	91.73 (2.86)	90.63 (3.52)	91.96 (2.80)
	○	●	○	●	24.90 (2.55)	23.83 (2.34)	24.47 (2.58)	23.05 (2.74)	26.86 (2.33)	90.93 (2.98)	89.16 (3.13)	90.49 (2.89)	89.02 (2.99)	92.03 (2.76)
	○	○	●	●	26.09 (3.11)	24.10 (2.89)	25.35 (2.67)	24.21 (2.54)	27.33 (2.25)	91.29 (3.07)	90.40 (3.10)	91.07 (2.71)	89.98 (3.21)	91.66 (2.70)
	○	●	●	●	26.23 (3.22)	24.66 (3.06)	25.54 (2.87)	24.76 (2.73)	27.84 (2.24)	92.15 (3.10)	91.25 (3.42)	91.80 (2.74)	90.35 (2.99)	92.95 (2.71)
T_2	●	○	○	○	23.20 (2.35)	23.28 (1.96)	22.33 (2.54)	20.53 (2.39)	23.56 (2.70)	88.31 (6.92)	88.57 (3.45)	87.76 (7.06)	85.10 (7.22)	88.47 (3.89)
	○	○	●	○	22.82 (2.26)	22.72 (1.75)	22.40 (2.18)	20.84 (2.14)	23.09 (1.89)	87.44 (6.55)	88.06 (2.71)	86.69 (6.48)	83.88 (6.56)	85.46 (2.83)
	○	○	○	●	23.16 (2.05)	23.24 (2.01)	22.05 (2.04)	20.88 (1.84)	23.71 (2.23)	87.06 (6.34)	87.69 (2.54)	86.70 (6.66)	83.66 (6.56)	88.15 (3.02)
	●	○	●	○	23.95 (2.53)	23.61 (1.84)	23.11 (2.35)	21.25 (2.34)	25.27 (2.30)	89.16 (6.72)	89.38 (3.03)	88.53 (6.70)	85.76 (6.93)	90.27 (3.38)
	●	○	○	●	24.45 (2.42)	24.42 (2.18)	23.53 (2.38)	21.81 (2.26)	26.22 (2.42)	90.07 (6.78)	90.52 (3.15)	89.74 (6.88)	87.04 (6.85)	91.87 (3.41)
	○	○	●	●	24.46 (2.44)	24.31 (1.98)	23.53 (2.24)	22.13 (2.24)	25.52 (1.98)	89.59 (6.54)	90.07 (2.55)	89.10 (6.58)	86.49 (6.58)	90.89 (2.97)
	●	○	●	●	24.89 (2.67)	24.49 (2.08)	24.00 (2.39)	22.32 (2.38)	26.06 (2.18)	90.67 (6.75)	90.92 (2.91)	90.21 (6.74)	87.51 (6.81)	92.60 (3.16)
T_{1ce}	●	○	○	○	22.70 (2.09)	22.88 (2.04)	21.92 (1.96)	22.11 (2.59)	24.15 (3.28)	87.15 (3.71)	85.62 (3.36)	86.98 (3.71)	85.74 (3.81)	87.28 (3.70)
	○	●	○	○	22.34 (2.19)	21.46 (1.79)	21.75 (1.89)	20.98 (2.03)	22.87 (2.08)	86.23 (2.92)	83.69 (2.62)	85.25 (2.81)	83.99 (2.86)	86.60 (2.96)
	○	○	○	●	22.12 (1.88)	21.48 (1.72)	21.58 (1.66)	20.68 (2.01)	22.72 (2.54)	84.85 (3.03)	83.26 (2.83)	83.95 (3.11)	80.71 (3.05)	85.00 (3.18)
	●	●	○	○	23.32 (2.47)	22.84 (2.12)	22.42 (1.88)	22.53 (2.53)	24.95 (3.10)	88.46 (3.26)	85.77 (3.00)	87.40 (3.40)	86.54 (3.47)	89.38 (3.34)
	●	○	○	●	22.99 (2.36)	22.78 (2.05)	22.18 (1.79)	22.54 (2.50)	24.62 (3.14)	87.98 (3.46)	85.25 (3.25)	87.27 (3.48)	85.72 (3.92)	89.17 (3.26)
	○	●	○	●	22.69 (2.13)	22.15 (1.88)	22.11 (1.81)	21.40 (2.06)	23.91 (2.61)	87.23 (2.85)	85.65 (2.66)	86.22 (2.74)	84.07 (2.85)	88.04 (2.89)
	●	●	○	●	23.30 (2.64)	22.95 (2.21)	22.49 (1.85)	22.76 (2.18)	25.00 (3.12)	88.69 (3.21)	86.14 (3.03)	87.82 (3.18)	86.06 (3.41)	90.15 (3.08)
T_{2f}	●	○	○	○	20.73 (2.27)	20.30 (2.01)	20.16 (1.93)	19.73 (1.49)	21.45 (2.22)	84.26 (3.08)	84.32 (2.77)	83.32 (3.15)	82.16 (3.29)	84.70 (3.42)
	○	●	○	○	22.63 (2.67)	22.48 (2.30)	21.39 (2.43)	20.96 (2.04)	22.98 (2.32)	85.34 (3.33)	85.42 (2.86)	85.50 (3.47)	84.28 (3.56)	85.55 (3.13)
	○	○	●	○	21.73 (2.33)	22.41 (2.14)	20.64 (1.96)	20.20 (1.74)	22.27 (1.95)	83.49 (3.22)	83.62 (2.86)	83.07 (3.31)	81.53 (3.08)	84.02 (2.97)
	●	●	○	○	23.22 (2.74)	23.91 (2.54)	21.81 (2.56)	21.86 (2.11)	24.51 (2.33)	87.81 (3.08)	87.73 (2.86)	87.01 (3.14)	85.77 (3.06)	88.46 (3.24)
	●	○	●	○	22.37 (2.35)	22.76 (2.12)	21.04 (2.16)	20.41 (1.70)	23.18 (2.05)	85.78 (3.01)	85.76 (2.74)	84.52 (3.20)	82.80 (3.04)	86.79 (3.23)
	○	●	●	○	23.27 (2.51)	23.92 (2.35)	22.00 (2.51)	21.92 (2.02)	24.29 (2.71)	87.68 (3.01)	87.55 (2.81)	86.72 (3.21)	85.13 (3.03)	88.09 (3.17)
	●	●	●	○	23.55 (2.63)	23.93 (2.32)	22.10 (2.63)	21.75 (1.85)	24.72 (2.70)	88.27 (2.99)	87.88 (2.75)	87.29 (3.12)	84.57 (2.83)	89.11 (3.18)

improved incrementally as the number of source (observed) modalities increased from one to three. For single-source scenarios, T1 synthesis was most effective when using T1ce or T2 as the source. For other targets, the best source modalities were FLAIR for T2 synthesis, T1 for T1ce synthesis, and T2 for FLAIR synthesis. Similarly, in two-source scenarios, the highest performance was achieved by the combinations of {T2, T1ce} for T1, {T1, FLAIR} for T2, and {T1, T2} for both T1ce and FLAIR. Using T1ce or FLAIR as the source did not always result in the highest scores, even though they are rich in tumor information. This is because the metrics are calculated on the whole image, rather than just tumor regions. This suggests that other structures and contrast outside the tumor are more heavily weighted when using other modalities as the source. This tendency was also observed in the baseline models.

Statistical analysis via t-tests revealed that our method achieved significant improvements over others in terms of PSNR across all tasks ($p < 0.05$). However, significant differences in SSIM were primarily observed in two-source tasks, with less improvement in single-source scenarios. To investigate this, we visualized results for three specific cases (T1 → T2, {T1, T2} → FLAIR, and {T1, T2, FLAIR} → T1ce), as shown in Fig. 4. In the single-source case, while our method produced noise-free images, they appeared slightly blurrier compared to the baselines, which likely affected the SSIM scores. Nevertheless, our model successfully reconstructed images similar to the ground truth without structural hallucinations. As the number of source modalities increased, the synthesized

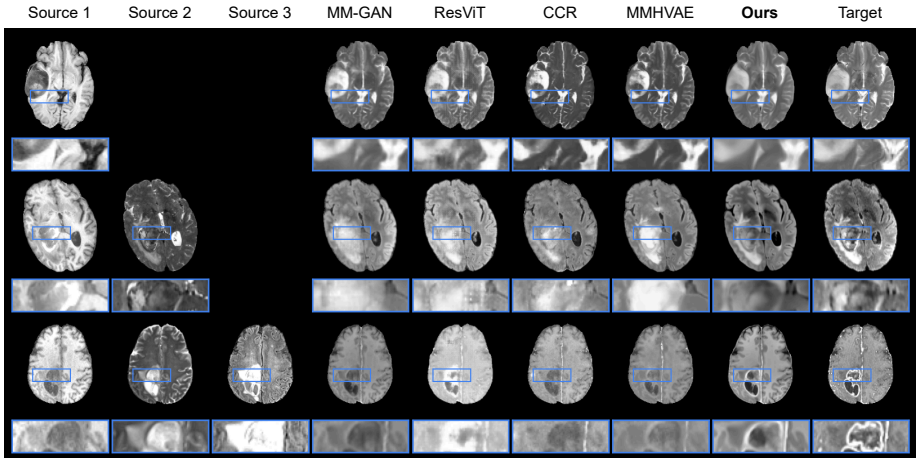

Fig. 4: Qualitative comparison of our method with baselines for synthesizing in various scenarios. **Top**: T1 → T2 synthesis. The bounding boxes show zoomed-in views of the brain tumor and ventricular regions. **Middle**: T1, T2 → FLAIR synthesis. The bounding boxes show zoomed-in views of the brain tumor and ventricular regions. **Bottom**: T1, T2, FLAIR → T1ce synthesis. The bounding boxes show zoomed-in views of the brain tumor region.

Table 2: Quantitative performance of MRI synthesis with baselines on the IXI. Mean and standard deviation values are presented. Blue indicates the best performance. Underline indicates statistical significance compared to all others ($p < 0.05$).

Target	Sources			PSNR (dB) ↑					SSIM (%) ↑				
	T_1	T_2	PD	MMGAN	ResViT	CCR	MMHVAE	**Ours**	MMGAN	ResViT	CCR	MMHVAE	**Ours**
T_1	○	●	○	23.84 (2.48)	24.00 (2.71)	21.64 (1.64)	23.76 (2.63)	24.64 (1.53)	90.10 (4.98)	90.29 (4.81)	89.84 (4.61)	88.28 (4.72)	90.76 (2.70)
	○	○	●	23.31 (2.22)	23.89 (2.57)	22.17 (1.66)	22.50 (2.14)	23.94 (1.09)	89.13 (4.55)	89.50 (4.93)	86.49 (3.73)	84.38 (4.14)	90.33 (1.46)
	○	●	●	24.24 (2.65)	24.79 (2.94)	21.68 (1.80)	24.09 (2.68)	25.93 (1.96)	91.16 (5.08)	91.33 (5.38)	89.97 (4.77)	87.77 (4.66)	91.72 (3.18)
T_2	●	○	○	21.20 (1.69)	20.82 (1.42)	20.11 (1.21)	21.86 (1.78)	22.78 (1.21)	88.45 (4.21)	88.79 (4.36)	86.69 (3.86)	88.28 (4.31)	89.51 (2.14)
	○	○	●	21.45 (1.39)	22.92 (1.44)	20.51 (1.06)	23.35 (1.43)	23.49 (0.99)	89.99 (2.79)	89.63 (3.05)	88.09 (2.92)	90.11 (2.97)	90.18 (1.40)
	●	○	●	22.18 (1.62)	23.70 (1.55)	22.27 (1.60)	24.24 (2.04)	25.38 (2.02)	91.16 (3.01)	91.22 (3.26)	91.36 (3.35)	91.58 (3.14)	91.80 (3.12)
PD	●	○	○	23.19 (1.84)	23.57 (1.90)	23.11 (1.57)	23.24 (1.78)	24.85 (1.10)	87.79 (3.93)	88.03 (4.16)	86.52 (3.65)	86.67 (3.64)	88.16 (2.48)
	○	●	○	26.07 (1.62)	26.39 (1.95)	25.66 (2.24)	26.03 (1.66)	26.64 (1.20)	91.04 (2.92)	90.61 (3.03)	89.47 (2.79)	90.63 (2.90)	91.64 (1.95)
	●	●	○	26.31 (1.60)	26.49 (2.18)	26.76 (2.23)	26.13 (1.93)	27.72 (1.69)	92.35 (2.94)	92.39 (3.30)	90.60 (2.96)	90.81 (2.91)	92.48 (2.59)

images became sharper. Unlike competing models that failed to accurately represent tumor regions, our method showed the highest fidelity to the ground truth in terms of overall contrast, structural detail. Specifically, T1ce synthesis, the most challenging task due to the difficulty of predicting enhancing tumor and tumor core regions, highlighted the robustness of our approach. While baselines failed to capture tumor details, often underestimating enhancing regions (MM-GAN, CCR, MMHVAE) or producing distortions (ResViT), our model achieved the most accurate and realistic depictions of the tumor.

Results on the IXI. We compared the missing modality synthesis results for all possible scenarios on the IXI, with the quantitative results summarized in Table 2. Unlike on the BraTS, which consists of tumor patient images, IXI contains normal brain scans. Consistent with the trends observed on the BraTS,

Table 3: Quantitative performance of MRI synthesis with different posterior sampling methods on the BraTS. Mean and standard deviation values are presented. Blue indicates the best performance. Cyan indicates the second best performance.

Target	Sources				PSNR (dB) ↑					SSIM (%) ↑				
	T_1	T_2	T_{1c}	T_{2f}	DPS	DDNM	PSLD	FlowChef	FlowDPS	DPS	DDNM	PSLD	FlowChef	FlowDPS
T_1	○	●	○	○	24.21 (2.01)	24.26 (1.93)	25.56 (2.39)	24.21 (1.99)	24.10 (1.89)	87.31 (2.64)	87.37 (2.62)	90.52 (3.12)	87.29 (2.65)	85.49 (2.97)
	○	○	●	○	25.32 (1.88)	25.32 (1.88)	25.56 (2.14)	25.31 (1.88)	25.43 (1.92)	88.34 (2.45)	88.36 (2.46)	88.17 (2.84)	88.35 (2.42)	88.16 (2.51)
	○	○	○	●	23.91 (2.39)	23.96 (2.37)	24.66 (2.61)	23.94 (2.35)	24.04 (2.21)	86.56 (2.70)	86.61 (2.76)	88.27 (3.06)	86.59 (2.75)	84.52 (6.29)
	○	●	●	○	27.01 (2.07)	26.99 (2.07)	27.15 (2.43)	26.94 (2.08)	27.11 (2.10)	91.86 (2.71)	91.83 (2.69)	91.96 (2.80)	91.85 (2.69)	91.90 (2.70)
	○	●	○	●	26.17 (2.37)	26.12 (2.36)	26.86 (2.33)	26.02 (2.31)	26.32 (2.18)	91.35 (2.66)	91.32 (2.68)	92.03 (2.76)	91.30 (2.69)	90.54 (2.70)
	○	○	●	●	26.68 (1.97)	26.71 (1.97)	27.33 (2.25)	26.73 (1.97)	26.84 (2.00)	91.05 (2.59)	91.08 (2.54)	91.66 (2.70)	91.10 (2.60)	90.75 (5.16)
	○	●	●	●	27.47 (2.11)	27.47 (2.08)	27.84 (2.24)	27.47 (2.07)	27.65 (2.05)	92.72 (2.67)	92.73 (2.64)	92.95 (2.71)	92.73 (2.65)	92.74 (2.68)
T_2	●	○	○	○	22.15 (2.36)	22.16 (2.36)	23.56 (2.70)	22.15 (2.41)	22.11 (2.27)	86.06 (3.39)	86.09 (3.41)	88.47 (3.89)	86.08 (3.44)	85.85 (3.40)
	○	○	●	○	21.45 (1.83)	21.58 (1.80)	23.09 (1.89)	21.50 (1.87)	21.49 (1.82)	83.73 (2.53)	83.83 (2.59)	85.46 (2.83)	83.76 (2.65)	83.62 (2.63)
	○	○	○	●	23.02 (2.07)	22.96 (2.14)	23.71 (2.23)	23.00 (2.07)	23.06 (1.97)	86.08 (2.82)	85.94 (2.88)	88.15 (3.02)	86.02 (2.77)	85.87 (2.83)
	●	○	●	○	23.65 (2.10)	23.68 (2.03)	25.27 (2.30)	23.72 (2.03)	23.70 (2.06)	88.40 (3.25)	88.46 (3.25)	90.27 (3.38)	88.43 (3.27)	88.38 (3.30)
	●	○	○	●	25.34 (2.09)	25.36 (2.08)	26.22 (2.42)	25.36 (2.07)	25.56 (2.24)	90.78 (3.27)	90.80 (3.27)	91.87 (3.41)	90.82 (3.27)	90.75 (3.33)
	○	○	●	●	24.62 (1.77)	24.72 (1.66)	25.52 (1.98)	24.60 (1.78)	24.92 (1.71)	89.21 (2.80)	89.26 (2.78)	90.89 (2.97)	89.21 (2.79)	89.21 (2.77)
	●	○	●	●	25.68 (1.93)	25.67 (1.87)	26.06 (2.18)	25.68 (1.90)	26.11 (1.87)	91.80 (3.11)	91.80 (3.09)	92.60 (3.16)	91.78 (3.12)	91.82 (3.11)
T_{1ce}	●	○	○	○	23.68 (2.98)	23.70 (2.95)	24.15 (3.28)	23.70 (2.95)	23.74 (2.51)	86.90 (3.15)	86.95 (3.19)	87.28 (3.70)	86.97 (3.19)	86.77 (3.18)
	○	●	○	○	22.10 (2.30)	22.14 (2.24)	22.87 (2.08)	22.09 (2.24)	22.03 (1.69)	84.89 (2.70)	84.95 (2.66)	86.60 (2.96)	84.88 (2.63)	81.18 (3.30)
	○	○	○	●	22.29 (2.42)	22.32 (2.40)	22.72 (2.54)	22.30 (2.39)	22.39 (1.91)	84.22 (2.84)	84.23 (2.84)	85.00 (3.18)	84.21 (2.87)	83.11 (2.97)
	●	●	○	○	24.62 (2.87)	24.60 (2.85)	24.95 (3.10)	24.60 (2.83)	24.50 (2.83)	88.98 (3.13)	88.97 (3.12)	89.38 (3.34)	88.97 (3.11)	88.93 (3.05)
	●	○	○	●	24.18 (2.81)	24.21 (2.82)	24.62 (3.14)	24.18 (2.80)	24.15 (2.79)	88.48 (2.99)	88.54 (2.97)	89.17 (3.26)	88.49 (2.98)	88.46 (2.98)
	○	●	○	●	23.42 (2.48)	23.44 (2.53)	23.91 (2.61)	23.44 (2.52)	23.51 (2.44)	87.57 (2.74)	87.60 (2.78)	88.04 (2.89)	87.60 (2.78)	87.70 (2.70)
	●	●	○	●	24.77 (2.82)	24.83 (2.83)	25.00 (3.12)	24.76 (2.82)	24.89 (2.74)	89.85 (2.97)	89.83 (2.97)	90.15 (3.08)	89.83 (2.97)	89.83 (2.98)
T_{2f}	●	○	○	○	20.29 (2.05)	20.32 (2.06)	21.45 (2.22)	20.24 (2.02)	20.25 (2.02)	83.77 (3.13)	83.75 (3.10)	84.70 (3.42)	83.78 (3.15)	83.69 (3.06)
	○	●	○	○	21.54 (2.05)	21.57 (2.11)	22.98 (2.32)	21.48 (2.09)	21.15 (2.14)	83.53 (2.87)	83.60 (2.91)	85.55 (3.13)	83.48 (2.86)	82.86 (2.69)
	○	○	●	○	21.24 (1.97)	21.26 (1.89)	22.27 (1.95)	21.23 (1.93)	21.12 (1.87)	83.77 (2.82)	83.87 (2.87)	84.02 (2.97)	83.81 (2.86)	83.88 (2.85)
	●	●	○	○	23.85 (2.68)	23.86 (2.58)	24.51 (2.33)	23.85 (2.63)	23.82 (2.67)	88.01 (3.18)	87.98 (3.16)	88.46 (3.24)	87.95 (3.19)	88.05 (3.15)
	●	○	●	○	22.06 (1.90)	22.12 (1.86)	23.18 (2.05)	22.05 (1.94)	22.07 (1.90)	86.09 (3.06)	86.14 (3.07)	86.79 (3.23)	86.09 (3.12)	86.12 (3.12)
	○	●	●	○	23.92 (2.14)	23.88 (2.28)	24.29 (2.71)	23.95 (2.18)	23.93 (2.24)	87.75 (3.07)	87.75 (3.04)	88.09 (3.17)	87.78 (3.06)	87.81 (3.05)
	●	●	●	○	24.27 (2.45)	24.22 (2.36)	24.72 (2.70)	24.34 (2.37)	24.50 (2.40)	88.88 (3.13)	88.87 (3.07)	89.11 (3.18)	88.87 (3.09)	88.98 (3.05)

both the baselines and the proposed method achieved higher performance when two source modalities were provided compared to a single-source scenario. In single-source tasks, the most effective source modalities were identified as T2 for T1 synthesis, PD for T2 synthesis, and T2 for PD synthesis. These results suggest that the synergy of multiple observed modalities significantly enhances the fidelity of the generated sequences, even in healthy brain structures. Our model also showed the highest performance in missing modality generation on IXI, proving the robustness of our method.

5.3 Results with Posterior Sampling methods

We evaluated various posterior sampling methods on the BraTS dataset, with the quantitative results presented in Table 3. Depending on the selected posterior sampling methods, Eq. 6 in Algorithm 1 is adapted accordingly (details in the supplement). Across most tasks, PSLD consistently achieved the highest performance. DPS, FlowChef, and FlowDPS employ gradient-based soft constraints; among these, FlowDPS yielded the best results, ranking second overall behind PSLD. DDNM, which utilizes a projection-based hard constraint to enforce measurement consistency, was next to FlowDPS in performance. PSLD demonstrated the highest image fidelity across most tasks by leveraging a hybrid approach that integrates both gradient-based and projection-based constraints. Specifically, PSLD achieved the best performance by computing the gradient and projection distance between the estimation and the measurement. This mechanism ensures that the synthesized images strictly adhere to the observed source

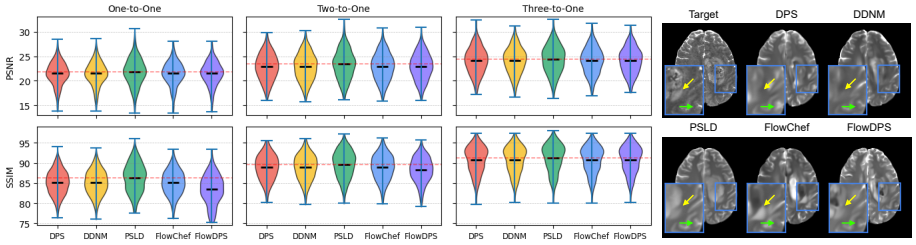

Fig. 5: Illustration of synthesis results on BraTS using various posterior sampling methods. Left: Quantitative performance plot relative to the number of source modalities. The red line shows the mean performance of PSLD, which achieved the best performance. Right: Qualitative comparison of T1 $\rightarrow$ T2 synthesis results with $N_{\text{EX}} = 1$. The yellow arrow indicates edema, while the green arrow indicates the cortex region.

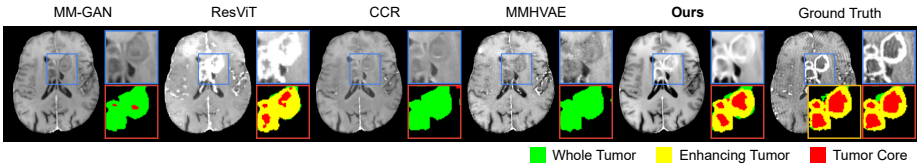

Fig. 6: Illustration of tumor segmentation results with synthesized T1ce (T1, T2, FLAIR $\rightarrow$ T1ce). The blue bounding boxes shows the tumor region.The red bounding boxes shows segmentation predictionm and the yellow bounding boxes shows the ground truth.

modalities while effectively preserving the generative prior to seamlessly reconstruct the missing modalities.

We also analyzed the performance of each method relative to the number of available source modalities, as illustrated in Fig. 5-left. PSLD consistently outperformed the other methods across one, two, three-source scenarios. The SSIM gap between PSLD and the baselines was most apparent when fewer source modalities were available, but it gradually diminished as the number of sources increased. The PSLD outperformed other methods in PSNR, though the differences were slight. Since PSNR is more affected by noise and contrast than structural metrics like SSIM, most models produced comparable results. This trend suggests that the synthesis process relies heavily on posterior sampling guidance when source information is scarce, whereas it depends more on vector field replacement (Eq. 5) when ample source modalities are provided. Fig. 5-right presents the qualitative results of $T_1 \rightarrow T_2$ synthesis for each method. The gradient-based methods (DPS, FlowChef, and FlowDPS) struggled to generate the tumor edema region and failed to accurately depict the cortex. Although DDNM successfully reconstructed the cortex, it exhibited weak contrast in the edema region. Conversely, PSLD reconstructed both regions with the highest fidelity to the ground truth. Based on these findings, we adopted PSLD as the default posterior sampling method for our framework.

Table 5: Ablation study on training strategies and sampling methods. MTM and MTO denote many-to-many and many-to-one sampling, respectively.

Task			$T_1 \rightarrow T_2$				$T_1, T_2 \rightarrow T_{2f}$			
Training	**MTM**	**MTO**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**
Full-modal only	✓		22.28 (2.49)	85.95 (3.66)	0.118 (0.028)	83.87	23.76 (2.50)	87.66 (3.18)	0.129 (0.024)	107.09
		✓	22.19 (2.63)	82.49 (5.81)	0.225 (0.027)	81.64	23.71 (2.34)	85.25 (5.11)	0.141 (0.023)	110.10
All scenarios	✓		22.26 (2.15)	83.69 (3.07)	0.123 (0.022)	65.75	23.86 (2.29)	87.95 (3.02)	0.122 (0.024)	95.67
		✓	23.56 (2.70)	88.47 (3.89)	0.101 (0.029)	54.15	24.51 (2.33)	88.46 (3.24)	0.109 (0.026)	87.96

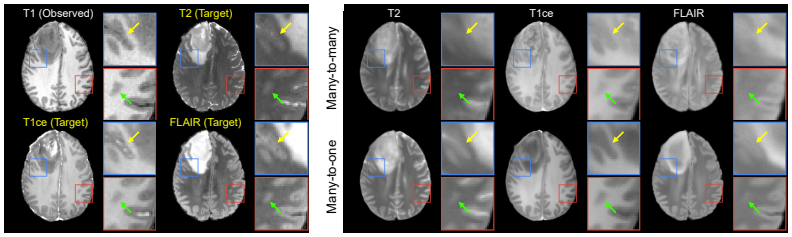

Fig. 7: Visual comparison between many-to-many and many-to-one sampling given an observed T1. The blue and red bounding boxes highlight different anatomical regions, while the yellow and green arrows indicate distinct cortex areas.

5.4 Impact on Tumor Segmentation

Table 4: Tumor segmentation performance with synthesized T1ce (T1, T2, FLAIR → T1ce).

Methods	Dice (%) ↑			
	WT	**ET**	**TC**	**Avg.**
MMGAN [43]	90.75 (4.99)	77.04 (15.14)	62.30 (21.55)	76.69 (10.98)
ResViT [11]	91.39 (5.78)	76.25 (16.18)	52.01 (24.26)	73.21 (12.65)
CCR [46]	85.75 (9.22)	53.44 (23.03)	38.11 (19.18)	59.10 (16.92)
MMHVAE [14]	84.25 (9.10)	50.81 (21.51)	36.99 (18.22)	57.35 (16.07)
Ours	91.53 (6.17)	79.97 (11.32)	63.79 (18.19)	78.43 (9.30)
Ground Truth	93.78 (4.20)	94.44 (5.34)	86.38 (17.98)	91.53 (7.29)

Preserving localized details is critical for brain tumor imaging. We therefore evaluated downstream tumor segmentation by feeding the synthesized T1ce images (T1, T2, and FLAIR → T1ce) into a pre-trained nnUNet [23]. The quantitative results are summarized in Table 4. Our method achieved the highest Dice scores for whole tumor (WT), enhancing tumor (ET), and tumor core (TC). Consistent with the baselines, ET and TC proved relatively more challenging to predict. Qualitative results are visualized in Fig 6. Models such as MM-GAN, CCR, and MMHVAE failed to capture the ET and TC regions, producing overly dark tumor contrasts. ResViT generated distorted tumor shapes. In contrast, our model reconstructed tumor details more accurately, yielding segmentation predictions closest to the ground truth. Moreover, as shown in supplementary Fig. B, reducing N_{EX} to 1 only slightly degrades downstream performance, indicating that the performance improvement is not solely due to expectation approximation sampling.

5.5 Ablation Study

We conducted an ablation study on our training and sampling strategies, with quantitative results summarized in Table 5. For training, we compared a full-

modal only (training on four modalities simultaneously from noise) against an all-scenarios (training on all combinations of two, three, or four modalities). For sampling, we evaluated many-to-many (MTM) sampling, which generates all missing modalities at once from the observed ones, versus many-to-one (MTO) sampling, which generates a single target modality. When trained on the full-modal only setup, the performance difference between MTM and MTO was marginal. This is likely because MTO requires masking other targets during sampling—a condition the model never encountered during full-modal training. However, when trained on All scenarios, MTO outperformed MTM in both image fidelity and quality. Visual comparisons between MTM and MTO are provided in Fig. 7. When synthesizing T2, T1ce, and FLAIR from a single T1 image, MTM failed to capture fine cortical details, whereas MTO preserved these subtle signals. These findings suggest that the iterative generation process in MTM is vulnerable to cross-modality interference, causing signal fading or hallucinations. The supplementary Fig. A shows this error propagation across sampling process.

5.6 Perception-Distortion Trade-off

Averaging multiple samples via expectation approximation sampling reduces noise and enhances the main signal. However, because the ground truth inherently contains textured noise, increasing N_{EX} improves image fidelity (lower distortion) at the cost of perceptual image quality. We observed the perception-distortion trade-off [5] during T1 $\rightarrow$ T2 synthesis (Fig. 8). At both NFE=10 and NFE=40, the highest perceptual quality is achieved at $N_{\text{EX}} = 1$. As N_{EX} increases, image fidelity increases while perceptual quality degrades. Furthermore, both quality and fidelity were notably higher at NFE=40 compared to NFE=10. This overall improvement is attributed to the measurement-guided vector field replacement (Eq. 5). Based on these findings, we empirically set NFE to 40 for optimal efficiency. Additional qualitative results across various NFE settings are provided in the supplementary material.

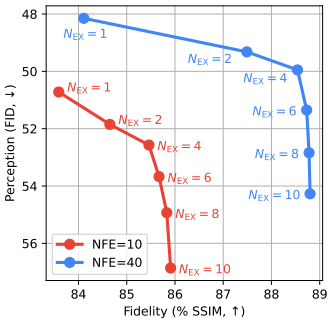

Fig. 8: The perception-distortion trade-off curve.

6 Conclusion

In this study, we demonstrate that complex architectures are unnecessary to build a generator capable of handling all missing modality scenarios. By simply training a flow-based model that iteratively samples data from noise, we showed that any missing modality can be effectively generated via posterior sampling. Furthermore, our approach achieved superior image fidelity compared to existing baselines. It also yielded the highest performance in downstream tumor segmentation, proving its capability to accurately synthesize not only global anatomical structures but also critical tumor regions with high fidelity to the ground truth.

Acknowledgement

We sincerely thank Professor **Hyunjin Park** for his valuable guidance, advice, and support throughout this research. This study was supported by the SMC-SKKU Future Convergence Research Program Grant (SMO125123). It was also supported by the AI Graduate School Support Program (Sungkyunkwan University) (RS-2019-II190421), the National Research Foundation of Korea (RS-2024-00408040), the ICT Creative Consilience Program Grant (IITP-2026-RS-2020-II201821), and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant (RS-2026-25528384, Resource-Intensive AI Technologies Based on Sustainable GPU Integrated Platforms), all funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

References

1. Aja-Fernández, S., Tristán-Vega, A.: A review on statistical noise models for magnetic resonance imaging. LPI, ETSI Telecomunicacion, Universidad de Valladolid, Spain, Tech. Rep (2013) 7
2. 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) 2, 3
3. Baid, U., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021) 8
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017) 8
5. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018) 8, 14
6. Brown, R.W., Cheng, Y.C.N., Haacke, E.M., Thompson, M.R., Venkatesan, R.: Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons (2014) 2
7. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: The Eleventh International Conference on Learning Representations (2023) 3, 4, 8
8. Chung, H., Lee, S., Ye, J.C.: Decomposed diffusion sampler for accelerating large-scale inverse problems. In: The Twelfth International Conference on Learning Representations (2024) 4
9. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. Advances in Neural Information Processing Systems **35**, 25683–25696 (2022) 4
10. Chung, H., Ye, J.C., Milanfar, P., Delbracio, M.: Prompt-tuning latent diffusion models for inverse problems. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=hrwIndai8e 4
11. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022) 2, 3, 8, 13

12. Daras, G., Chung, H., Lai, C.H., Mitsufuji, Y., Ye, J.C., Milanfar, P., Dimakis, A.G., Delbracio, M.: A survey on diffusion models for inverse problems. arXiv preprint arXiv:2410.00083 (2024) 3
13. De Coene, B., Hajnal, J.V., Gatehouse, P., Longmore, D.B., White, S.J., Oatridge, A., Pennock, J., Young, I., Bydder, G.: Mr of the brain using fluid-attenuated inversion recovery (flair) pulse sequences. American journal of neuroradiology **13**(6), 1555–1564 (1992) 1
14. Dorent, R., Haouchine, N., Golby, A., Frisken, S., Kapur, T., Wells, W.: Unified cross-modal medical image synthesis with hierarchical mixture of product-of-experts. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2, 3, 8, 13
15. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011) 4
16. Filippi, M., Rocca, M.A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., Rovira, A., Sastre-Garriga, J., Tintorè, M., Frederiksen, J.L., et al.: Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. The Lancet Neurology **15**(3), 292–303 (2016) 1
17. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016) 7
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf 3
19. Gudbjartsson, H., Patz, S.: The rician distribution of noisy mri data. Magnetic resonance in medicine **34**(6), 910–914 (1995) 7
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 8
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 3
23. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 13
24. 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) 3
25. Katti, G., Ara, S.A., Shireen, A.: Magnetic resonance imaging (mri)–a review. Int. J. Dent. Clin **3**(1), 65–70 (2011) 1
26. Kim, J., Kim, B.S., Ye, J.C.: Flowdps : Flow-driven posterior sampling for inverse problems. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12328–12337 (October 2025) 4, 8
27. Kim, J., Park, H.: Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7604–7613 (2024) 2, 3

28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
29. Kong, L., Lian, C., Huang, D., li, z., Hu, Y., Zhou, Q.: Breaking the dilemma of medical image-to-image translation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 1964–1978. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/0f2818101a7ac4b96ceeba38de4b934c-Paper.pdf 2, 3
30. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging **38**(11), 2556–2568 (2019) 2
31. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t 3
32. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z 3
33. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022) 6
34. Martin, S.T., Gagneux, A., Hagemann, P., Steidl, G.: Pnp-flow: Plug-and-play image restoration with flow matching. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=5AtHrq3B5R 6
35. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014) 1, 2, 8
36. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023) 2, 3
37. 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) 8
38. Patel, M., Wen, S., Metaxas, D.N., Yang, Y.: Flowchef: Steering of rectified flow models for controlled generations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15308–15318 (October 2025) 4, 8
39. Rassmann, S., Kügler, D., Ewert, C., Reuter, M.: Regression is all you need for medical image translation. arXiv preprint arXiv:2505.02048 (2025) 5, 7
40. Roemer, P.B., Edelstein, W.A., Hayes, C.E., Souza, S.P., Mueller, O.M.: The nmr phased array. Magnetic resonance in medicine **16**(2), 192–225 (1990) 7
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 8
42. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. Advances in Neural Information Processing Systems **36**, 49960–49990 (2023) 4, 8

43. Sharma, A., Hamarneh, G.: Missing mri pulse sequence synthesis using multi-modal generative adversarial network. IEEE transactions on medical imaging **39**(4), 1170–1183 (2019) 2, 3, 8, 13
44. 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), https://openreview.net/forum?id=PxTIG12RRHS 3
45. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. In: The Eleventh International Conference on Learning Representations (2023) 3, 4, 8
46. Xiong, H., Wang, Y., Shen, Z., Sun, K., Fang, Y., Chen, Y., Shen, D., Wang, Q.: Learning contrast and content representations for synthesizing magnetic resonance image of arbitrary contrast. Medical Image Analysis p. 103635 (2025) 2, 3, 8, 13
47. 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) 2, 3
48. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014) 7
49. Zhang, B., Chu, W., Berner, J., Meng, C., Anandkumar, A., Song, Y.: Improving diffusion inverse problem solving with decoupled noise annealing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20895–20905 (2025) 4
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 8