Patient-Adaptive Modality Attention for Multimodal Brain Tumor Synthesis via Latent Diffusion

Kyuwon Park[1,3], Xiaofeng Liu[2], Fangxu Xing[3], Helen A. Shih[3], Georges El Fakhri[2], Jae Youn Hwang[1], and Jonghye Woo[3]

[1] DGIST, Daegu, South Korea
[2] Yale University, New Haven, CT, USA
[3] Harvard Medical School and Massachusetts General Brigham, Boston, MA, USA
kw040131@dgist.ac.kr

Abstract. Data scarcity and class imbalance in brain tumor datasets remain critical bottlenecks for training robust diagnostic AI models. Synthesizing realistic tumors across multiple MRI modalities is challenging because sequences differ in sensitivity to pathological subregions, modality quality varies per subject, and strict inter-modal consistency must be maintained. We propose a subject-specific multimodal latent diffusion framework for high-fidelity brain tumor synthesis via *Disease-Aware Modality AttentioN* (DAMAN). Separate modality-specific 3D VQ-GANs encode each MRI contrast into a compact latent space; our framework dynamically computes global quality-aware gating weights and voxel-level lesion-specific spatial attention to adaptively fuse complementary cross-modal features for each patient. A hard-inpainting loop at every reverse diffusion step enforces strict background preservation. Evaluated on the LUMIERE longitudinal glioblastoma dataset with four MRI contrasts under a patient-level train/test split, our framework achieves PSNR of 20.27 dB and SSIM of 0.919 on the target modality (T1-post), outperforming single-modal baselines. An independent, harmonization-free, whole-image evaluation across all four modalities further confirms these gains and shows that DAMAN reaches a 100% valid generation rate, compared with 50–56% for the baselines.

Keywords: Brain Tumor Synthesis · Adaptive Attention · Multimodal Fusion · Multi-contrast MRI · Latent Diffusion

1 Introduction

Multi-contrast brain MRI is critical for diagnosing and monitoring intracranial tumors such as glioblastoma [7]. Radiologists routinely interpret complementary sequences—T1-pre, T1-post, T2, and FLAIR—to delineate distinct sub-regional structures [11]. Despite rapid progress in deep learning for tumor detection and segmentation [8], robust model development remains limited by data scarcity and severe class imbalance, particularly for atypical lesions [14,2,5]. Realistic,

controllable synthetic tumor generation is therefore a principled approach to augment training data while preserving anatomical plausibility.

Related work. GAN-based [3] and VQ-VAE-style [9] methods demonstrated high-fidelity 2D medical image synthesis but scale poorly to 3D volumes. Diffusion models [4,10] address this via iterative denoising in a compressed latent space. For brain tumor synthesis, prior work has adapted latent diffusion to 3D mask-conditioned multi-contrast generation using fixed equal-weight fusion [13], which is brittle to modality-specific quality variations. In missing-modality imputation and multi-site harmonization, methods such as SynthSeg [1] and HACA3 [16] address cross-contrast or site-induced inconsistencies but are not designed for pathological inpainting. Unified frameworks [6] further address multi-site harmonization yet treat all modalities symmetrically. Our work differs in explicitly modeling *per-subject, per-voxel* modality reliability within a mask-conditioned latent diffusion framework: MRI quality varies per acquisition due to motion, field inhomogeneity, or absent sequences, making uniform fusion suboptimal and motivating patient-adaptive conditioning.

We propose a two-stage pipeline: a modality-specific 3D VQ-GAN compresses multi-contrast MRI into a compact latent space, and our *Disease-Aware Modality AttentioN* (DAMAN) framework performs mask-conditioned latent diffusion with adaptive modality fusion. Our primary contributions are: (1) to our knowledge, the first 3D latent diffusion framework with *patient-adaptive* multimodal fusion for brain tumor inpainting, with anatomical preservation enforced via latent hard-inpainting; (2) a novel adaptive fusion module with global quality-aware gating and lesion-specific spatial attention for patient-specific modality weighting; and (3) a systematic evaluation across four MRI contrasts under a patient-level train/test split demonstrating consistent synthesis fidelity and 100% valid generation rate across all modalities.

2 Methodology

To synthesize highly realistic and anatomically plausible brain tumors while maintaining cross-modal consistency, we propose a two-stage mask-conditioned latent diffusion framework (Fig. 1).

2.1 Stage 1: Modality-Specific 3D VQ-GAN

Directly modeling the diffusion process in a high-dimensional 3D voxel space is memory-intensive and prone to unstable training. We therefore train a separate VQ-GAN [3] for each MRI contrast, clipping intensities to the 1^{st}–99^{th} percentile to suppress outlier peaks. For a normalized volume $\tilde{x} \in \mathbb{R}^{1\times D\times H\times W}$, encoder E maps it to $z = E(\tilde{x})$, which is then element-wise quantized against a learnable codebook $\mathcal{C} = \{c_k\}_{k=1}^{K} \subset \mathbb{R}^C$:

$$z_q(u) = c_{k^*(u)}, \qquad k^*(u) = \arg\min_{k\in\{1,\dots,K\}} \|z(u) - c_k\|_2^2, \tag{1}$$

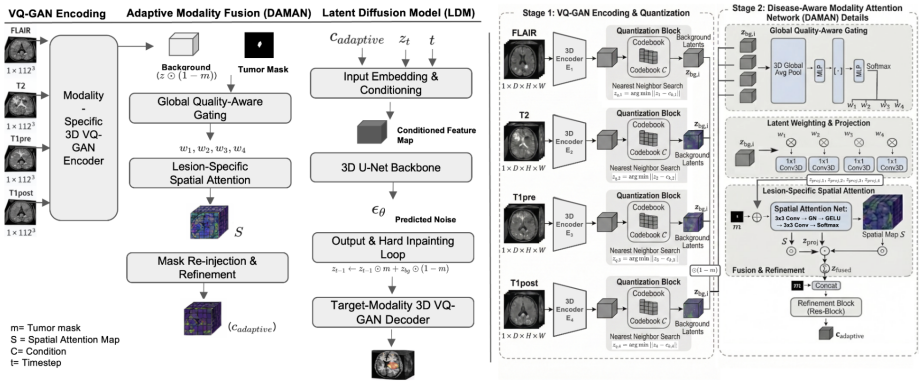

Fig. 1. Overview of the proposed framework. (*Left*) The pipeline comprises VQ-GAN encoding, DAMAN-based adaptive modality fusion, and LDM-based target synthesis with hard-inpainting. (*Right*) Detailed architecture: Stage 1 shows per-modality VQ-GAN quantization of each contrast into background latents; Stage 2 shows DAMAN internals—global quality-aware gating, latent weighting and projection via 1×1 Conv3D, lesion-specific spatial attention, and mask-conditioned fusion with a residual refinement block to produce c_{adaptive}.

where u is the latent spatial location. Decoder D reconstructs the volume from z_q. After training with reconstruction and commitment losses, the encoder and decoder are frozen, providing a compact, expressive latent space for diffusion [12].

Implementation details. Each modality-specific VQ-GAN uses a 3-level encoder/decoder (base channels = 64, codebook size $K = 1{,}024$, embedding dimension $C = 256$). Inputs are resampled to 112^3 voxels; latents are $28^3 \times 256$ channels (compression ratio 4×). A per-modality scale factor (reciprocal of latent standard deviation, ranging from 0.12 to 0.19) normalizes the latent space for stable diffusion training.

2.2 Stage 2: DAMAN and Mask-Conditioned Latent Diffusion

The generative process is formulated as inpainting within the VQ-GAN latent space. Let $m \in \{0,1\}^{1\times d\times h\times w}$ be the binary tumor mask downsampled to latent resolution, and $z_{bg,i} = z_{0,i} \odot (1-m)$ be the healthy background latent for modality i.

Why "patient-adaptive"? MRI quality varies per acquisition: motion, field inhomogeneity, or absent sequences produce modality-specific degradation that is unique to each patient and cannot be anticipated at training time. DAMAN addresses this by computing data-driven weights *at inference time* from the actual background latents, rather than using fixed fusion weights—making the fusion *patient-adaptive* by design.

Global quality-aware gating. DAMAN first applies 3D adaptive average pooling to each background latent to extract a global descriptor $v_i \in \mathbb{R}^C$. An MLP predicts unnormalized modality scores $a \in \mathbb{R}^M$, converted to normalized global weights $w_i = \exp(a_i)/\sum_{j=1}^{M} \exp(a_j)$. The scaled background latent for modality i is then $\tilde{z}_i = w_i z_{bg,i}$. The learned weights are intended to down-weight missing or less informative modalities.

Lesion-aware spatial attention. To capture the spatial heterogeneity of tumors (e.g., enhancement in T1-post versus edema in FLAIR), a 3×3 convolutional block generates a spatial attention map $S \in \mathbb{R}^{M \times d \times h \times w}$:

$$S = \mathrm{Softmax}(\mathrm{Conv3D}_{3\times3}(\sigma(\mathrm{GN}(\mathrm{Conv3D}_{3\times3}([\tilde{z}_1; \ldots; \tilde{z}_M; m]))))), \tag{2}$$

where σ is Gaussian Error Linear Unit and GN is Group Normalization. The softmax is applied along the modality dimension at each voxel, so that $\sum_{i=1}^{M} S_i(u) = 1$ for each latent voxel u. The spatially fused latent $z_{\mathrm{fused}} = \sum_i S_i \odot z_{\mathrm{proj},i}$ aggregates cross-modal features, where $z_{\mathrm{proj},i}$ is a linear projection of $\tilde{z}_i$ so that the global gating weights are propagated into the final fusion. A final refinement block re-injects the binary mask: $z_{\mathrm{out}} = \mathrm{Refine}([z_{\mathrm{fused}}; m]) + z_{\mathrm{fused}}$, and the adaptive condition is $c_{\mathrm{adaptive}} = [z_{\mathrm{out}}; m]$.

Diffusion objective. A 3D U-Net ϵ_θ (base channels = 64, two downsampling levels; $T = 1{,}000$ steps, linear β-schedule 10^{-4}–0.02; AdamW, lr $= 10^{-4}$, batch = 8, 300 epochs) is optimized to predict added Gaussian noise:

$$\mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_{\mathrm{adaptive}})\|_2^2 \right]. \tag{3}$$

2.3 Inference: Latent Hard-Inpainting

At each reverse diffusion step t, non-tumor regions are deterministically restored from the target background latent:

$$z_{t-1} \leftarrow z_{t-1} \odot m + z_{bg,\mathrm{target}} \odot (1 - m), \tag{4}$$

ensuring that background tissue is identical to the real scan throughout sampling and that no test-time information about the tumor region is introduced from the ground truth.

3 Experiments and Results

3.1 Experimental Setup

Dataset and split. All experiments use the LUMIERE dataset [11], comprising fully paired longitudinal multi-contrast MRI (T1-pre, T1-post, T2, FLAIR) with automated glioblastoma segmentation masks. We adopt a patient-level 80/20 train/test split (73 train / 18 test patients, seed = 42), ensuring no subject-level

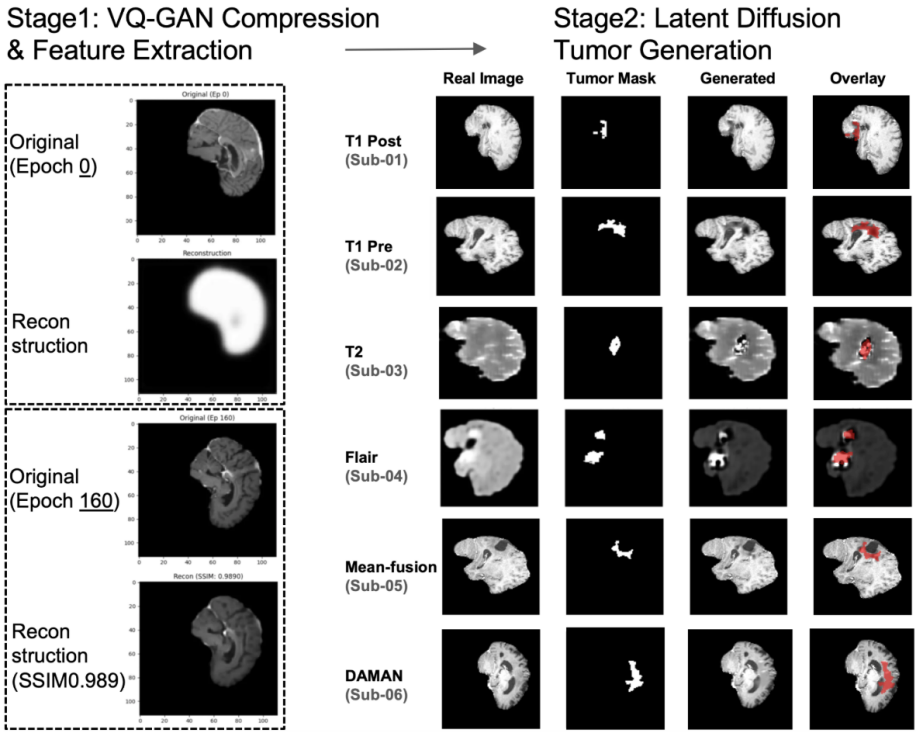

Fig. 2. Visual overview of the generation pipeline. (*Left*) Stage 1: VQ-GAN training progression from epoch 0 (blurry) to epoch 160 (high-fidelity reconstruction). (*Right*) Stage 2: synthesis results showing T1-post, T1-pre, T2, and FLAIR under single-modality conditioning, and Mean Fusion and DAMAN on T1-post; each row shows the real image, tumor mask, generated output, and overlay.

data leakage between training and evaluation. Both the VQ-GAN and all diffusion models are trained exclusively on the 73 training-split patients; per-modality scale factors are computed from training latents only. Scans with empty segmentation masks are excluded; all reported metrics are evaluated on the held-out test split, yielding up to 125 valid scans per modality after excluding samples with empty tumor masks.

Evaluation metrics. Because hard-inpainting perfectly preserves the healthy background, standard whole-image evaluation would artificially inflate scores. All primary tumor-quality metrics are therefore computed *exclusively within the tumor mask*: **PSNR** and **L1 error** for voxel-level fidelity; **SSIM** on three orthogonal center slices of the volume for structural realism. All reported metrics are evaluated on the held-out test split under a patient-level split with no subject overlap between training and evaluation. We primarily report quantitative results on T1-post, as it provides the most clinically informative visualization

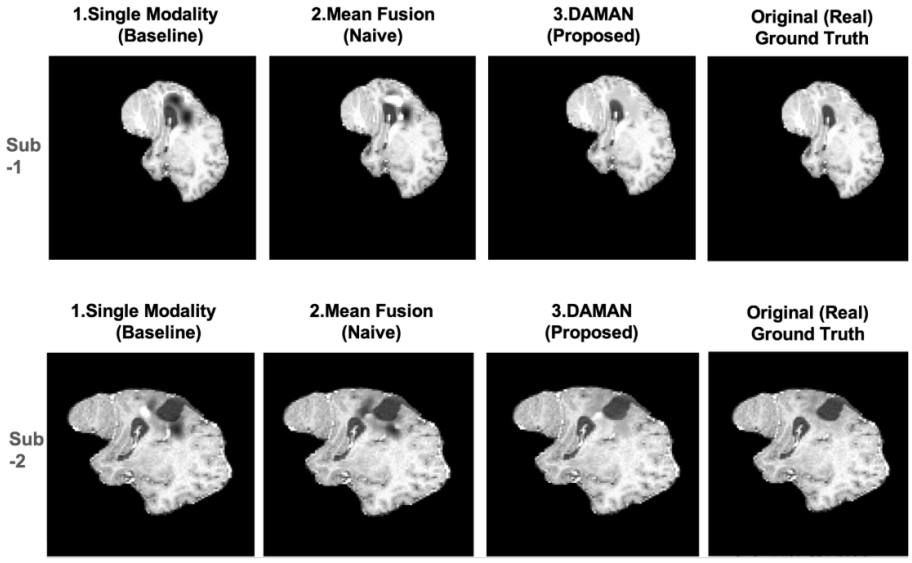

Fig. 3. Qualitative comparison for T1-post tumor synthesis. All three methods produce anatomically plausible tumors with preserved healthy background due to hard-inpainting; subtle differences in internal texture sharpness are visible between Single Modality, Mean Fusion, and DAMAN, consistent with their close quantitative performance in Table 2.

of contrast-enhancing tumor regions. Other modalities (T1-pre, T2, FLAIR) are included for completeness and cross-modality validation.

Baselines. We compare three configurations sharing the same VQ-GAN and U-Net backbone, differing only in the conditioning mechanism: (1) **Single Modality**—standard single-contrast latent diffusion without any cross-modal input; (2) **Mean Fusion**—naive equal-weight averaging of all four modality latents as the condition; (3) **DAMAN**—our proposed method.

3.2 Single-Modal Generation Across Modalities

Table 1 reports single-modal synthesis performance across all four contrasts. All four modalities achieve comparable PSNR (18.3–19.6 dB) and SSIM (≈0.92), demonstrating consistent synthesis quality. The slightly lower PSNR for T2 and FLAIR relative to T1-pre reflects the broader intensity range of these edema-sensitive contrasts within the tumor mask, motivating cross-modal guidance.

3.3 Fusion Strategy Comparison on T1-post

Table 2 compares the three conditioning strategies on T1-post synthesis. Single Modality achieves 19.38 dB PSNR and SSIM of 0.918, lacking cross-modal tex-

Table 1. Single-modal synthesis performance across all four MRI contrasts (LUMIERE, held-out test split). Metrics are computed within the tumor mask region.

Target Modality	PSNR (dB)↑	SSIM↑	L1 Error↓
T1-post	19.38 ± 6.76	0.918 ± 0.016	0.107 ± 0.039
T1-pre	19.55 ± 2.98	0.926 ± 0.018	0.106 ± 0.040
T2	18.56 ± 4.47	0.925 ± 0.018	0.123 ± 0.064
FLAIR	18.33 ± 4.94	0.923 ± 0.019	0.129 ± 0.054

Table 2. Quantitative evaluation on T1-post (held-out test split). Metrics computed within the tumor mask. Bold indicates the best value per column; tied values are both bolded. Generation reliability (valid/total) is reported in the text.

Method	PSNR (dB)↑	SSIM↑	L1 Error↓
Single Modality (Baseline)	19.38 ± 6.76	0.918 ± 0.016	0.107 ± 0.039
Mean Fusion (Naive)	20.19 ± 9.29	$\mathbf{0.919 \pm 0.016}$	$\mathbf{0.105 \pm 0.041}$
DAMAN (Proposed)	$\mathbf{20.27 \pm 9.68}$	$\mathbf{0.919 \pm 0.016}$	$\mathbf{0.105 \pm 0.041}$

ture cues. Both multimodal fusion strategies consistently improve over single-modality synthesis: Mean Fusion raises PSNR to 20.19 dB and DAMAN to 20.27 dB, both achieving SSIM of 0.919 and L1 of 0.105. The fidelity gain comes primarily from multimodal fusion itself, with DAMAN's adaptive weighting providing a marginal additional PSNR improvement. A further difference emerges in generation reliability: Single Modality and Mean Fusion yield a valid (non-divergent) output for only 67/125 cases (53.6%), whereas DAMAN produces a valid output for all 125/125 cases (100%), indicating that the quality-aware gating substantially stabilizes the reverse diffusion process relative to fixed and uniform conditioning.

3.4 Independent Verification Across All Modalities

To verify that the conclusions above are not an artifact of the tumor-mask evaluation protocol or specific to T1-post, we repeat the full comparison using a stricter setting: whole-image metrics computed without any test-time intensity adjustment, across all four target modalities (n=126 per modality). Table 3 reports PSNR results; absolute values differ from Table 2 as metrics here are computed over the full volume rather than within the tumor mask. DAMAN achieves the best PSNR in three of four targets (FLAIR, T2, T1-post) and is 0.11 dB behind the best on T1-pre, while producing valid output for 100% of test cases (126/126) across every modality. Single Modality and Mean Fusion yield valid outputs for only 50–56% of cases across all targets, confirming that DAMAN's generation reliability generalizes beyond T1-post.

Figs. 2 and 3 illustrate representative outputs from the full pipeline. Consistent with the quantitative results in Table 2, all three conditioning strategies produce anatomically plausible tumors with well-preserved healthy background.

Table 3. Independent verification: whole-image PSNR (dB), no test-time intensity adjustment, across all four target modalities (n=126, held-out test split). Bold indicates the best value per row. DAMAN valid: 126/126 (100%) for all targets; Single and Mean: 50–56%.

Target	Single	Mean Fusion	DAMAN
FLAIR	21.10 ± 1.58	21.02 ± 1.61	$\mathbf{21.23 \pm 1.44}$
T2	18.46 ± 1.57	18.44 ± 1.66	$\mathbf{18.53 \pm 1.57}$
T1-pre	19.33 ± 1.81	$\mathbf{19.61 \pm 1.91}$	19.50 ± 1.90
T1-post	10.47 ± 1.00	10.54 ± 0.99	$\mathbf{10.58 \pm 0.97}$

Single Modality synthesis shows slightly less internal textural detail without cross-modal cues [15], while Mean Fusion and DAMAN both leverage complementary multi-contrast information to produce comparable, heterogeneous tumor textures.

4 Conclusion

We presented DAMAN, a subject-specific multimodal latent diffusion framework for high-fidelity brain tumor synthesis. By dynamically computing global quality-aware gating weights and voxel-level spatial attention within a 3D VQ-GAN latent space, DAMAN adaptively integrates complementary multi-contrast features for each patient, while a latent hard-inpainting loop strictly preserves healthy background tissue. Under a strict patient-level train/test split, DAMAN achieves the best PSNR (20.27 dB) among all fusion strategies on T1-post, with both multimodal fusion strategies substantially improving fidelity over single-modality synthesis. A separate, harmonization-free, whole-image evaluation across all four MRI contrasts corroborates these fidelity gains and shows that DAMAN reaches a 100% valid generation rate (vs. 50–56% for the baselines), indicating that the quality-aware gating and attention mechanisms also stabilize the underlying generative process. Limitations include reliance on pre-defined tumor masks, fully paired multi-contrast training data, and a small training cohort (73 patients) which may constrain the degree of patient-specificity learned by the attention module. Future work will explore joint anatomy-and-mask synthesis, larger multi-site datasets, and attention diversity regularization to better realize the intended cross-modal specialization.

Disclosure of Interests

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

Acknowledgements

This work is partly supported by NIH R01CA290745 and R21EB034911.

References

1. Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., et al.: Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical image analysis **86**, 102789 (2023)
2. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering **5**(6), 493–497 (2021)
3. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
5. Jordon, J., Yoon, J., Van Der Schaar, M.: Pate-gan: Generating synthetic data with differential privacy guarantees. In: International conference on learning representations (2018)
6. Liu, H., Fan, Y., Cui, C., Su, Z., Kuo, L., Oguz, I., Dawant, B.M.: One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. IEEE Transactions on Medical Imaging **42**(8), 2418–2430 (2023)
7. Liu, X., Shih, H.A., Xing, F., Santarnecchi, E., El Fakhri, G., Woo, J.: Incremental learning for heterogeneous structure segmentation in brain tumor mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2023)
8. Liu, X., Xing, F., El Fakhri, G., Woo, J.: Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation. Medical image analysis **83**, 102641 (2023)
9. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
11. Suter, Y., Knecht, U., Valenzuela, W., Notter, M., Hewer, E., Schucht, P., Wiest, R., Reyes, M.: The lumiere dataset: Longitudinal glioblastoma mri with expert rano evaluation. Scientific data **9**(1), 768 (2022)
12. Tang, Y., Yang, D., Roth, H.R., et al.: High-resolution 3d medical image synthesis with vector-quantized latent space. arXiv preprint arXiv:2212.05938 (2022)
13. Truong, N.C., et al.: Synthesizing 3D multicontrast brain tumor MRIs using tumor mask conditioning. In: Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications. vol. 12931, pp. 116–120. SPIE (2024)
14. Xing, F., Liu, X., Kuo, C.C.J., El Fakhri, G., Woo, J.: Brain mr atlas construction using symmetric deep neural inpainting. IEEE journal of biomedical and health informatics **26**(7), 3185–3196 (2022)
15. Zhou, Y., Li, Z., He, N., et al.: Latent correlation representation learning for multimodal brain tumor segmentation. In: MICCAI (2021)
16. Zuo, L., Liu, Y., Xue, Y., Dewey, B.E., Remedios, S.W., Hays, S.P., Bilgel, M., Mowry, E.M., Newsome, S.D., Calabresi, P.A., Resnick, S.M., Prince, J.L., Carass, A.: HACA3: A unified approach for multi-site MR image harmonization. Computerized Medical Imaging and Graphics **109**, 102285 (2023)