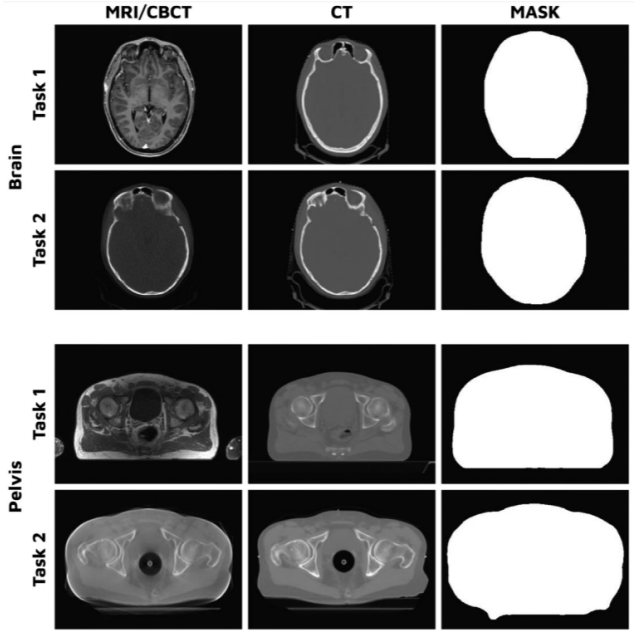

Figure 1. Example images for all tasks and anatomies in the synthRAD2023 dataset. The top row shows images for task 1 brain, the middle-top for task 1 pelvis, the middle-bottom for task 2 brain, and the bottom for task 2 pelvis. The first column shows the input images for each task: MRI (task 1) or CBCT (task 2); the second column is the ground truth CT, and the third column is the associated dilated body outline.

A. Datasets

SynthRAD2023 SynthRAD2023 is a synthetic radiology dataset designed for cross-modal brain image reconstruction tasks. It contains paired Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) volumes with various degrees of sparsity and anatomical variability. The dataset simulates realistic clinical conditions by incorporating sparse CT slices and multiple MRI modalities, facilitating the evaluation of reconstruction methods under challenging sparse input scenarios. Figure 1 shows example images from the dataset. The images are reproduced from the original paper.

BraTS The Brain Tumor Segmentation (BraTS) [3] dataset is a widely used benchmark for brain tumor imaging studies. It includes multi-modal MRI scans (such as T1, T1c, T2, and FLAIR) from patients with gliomas, along with expert annotations. BraTS provides high-resolution volumetric MRI data with diverse tumor appearances, making it suitable for evaluating cross-modal synthesis and reconstruction algorithms in neuro-oncology. In our work, we utilize BraTS to validate the robustness of our method on real clinical MRI data with complex pathological structures. Figure 5 shows example images, reproduced from the original paper.

B. Implementation Details

Due to space limitations in the main paper, we present detailed training procedures in this section, with particular emphasis on slice selection and training logic.

We begin by training a BBDM [2] using paired CT and MRI data. This supervised training phase is relatively straightforward, with no additional data alignment or augmentation required beyond standard preprocessing.

Next, we construct a brain knowledge base by organizing the training data according to subject IDs. For datasets such as BraTS and SynthRAD2023, the original directory structures are inherently organized by subject, naturally forming an efficient retrieval database without the need for restructuring.

To enable retrieval-based synthesis, we fine-tune a retrieval model solely on the CT modality. We explored the potential of using all four modalities in BraTS—namely T1, T1ce (T1 with contrast enhancement), T2, and FLAIR—to improve generalization of the retrieval model. However, empirical results show that this multimodal input often degrades performance. We hypothesize this is due to the monotonous nature of grayscale medical images, where increasing the number of input channels introduces noise rather than useful diversity.

For training the ControlNet, we follow a retrieval-augmented setting. Given a source CT image y, we retrieve another CT image r using the retrieval model. Our objective is to reconstruct the corresponding MRI of r, using y as the input to the frozen BBDM and r as the conditioning input to ControlNet. Only the ControlNet requires training in this phase. This significantly reduces training time compared to the original BBDM, as the parameter size of ControlNet is approximately half that of BBDM. The training process mimics a multi-agent setup: ControlNet is entirely dependent on the quality of the retrieval results. If the retrieval model performs poorly or the knowledge base lacks similar examples, ControlNet training becomes unstable. To mitigate this issue, we optionally apply spherical linear interpolation (slerp) between the retrieved image r and the input y, generating additional intermediate samples to enhance training stability.

Joint training of BBDM and ControlNet presents further challenges. In particular, we observe significant loss fluctuations, especially when the retrieved image r differs substantially from the input y. In such cases, the reconstructed MRI often lacks fine anatomical details. We refer to related work such as EBDM [1], and conclude that the bottleneck primarily lies in data availability and retrieval quality.

In most cases, when y and r are reasonably similar, the resulting output exhibits the grayscale style of y and the structural content of r. Occasionally, the output becomes

nearly indistinguishable from r, with minimal perceptual differences.

As discussed in the main paper, these limitations primarily stem from the lack of sufficiently similar samples in the retrieval database. Nevertheless, our method consistently outperforms baselines in terms of structural fidelity and style preservation, as evidenced by SSIM and PSNR scores.

C. Unitized vs. Non-Unitized Objectives

C.1. Theoretical Analysis of Objective Functions

Setup. Let x_0 denote the reference image and y the conditional image. Define

$$r = \|y - x_0\| \geq 0, \quad u = \frac{y - x_0}{\|y - x_0\|}, \quad y - x_0 = ru. \tag{1}$$

Fix timestep t and write m_t, δ_t; set $s = \sqrt{\delta_t}$. Let $\epsilon \sim \mathcal{N}(0, I)$ be independent Gaussian noise. Denote the model prediction as f.

$$\text{Raw objective: } L_{\text{raw}}(f) = \mathbb{E}\big[\|m_t ru + s\epsilon - f\|_2^2\big], \tag{2}$$

$$\text{Unitized objective: } L_{\text{unit}}(f) = \mathbb{E}\big[\|m_t u + s\epsilon - f\|_2^2\big]. \tag{3}$$

Proposition 1. The minimizers satisfy

$$f^\star_{\text{raw}} = m_t \mathbb{E}[ru \mid \mathcal{I}], \quad f^\star_{\text{unit}} = m_t \mathbb{E}[u \mid \mathcal{I}]. \tag{4}$$

In general $f^\star_{\text{raw}} \neq f^\star_{\text{unit}}$.

Proof. For ℓ_2 loss, the minimizer is the conditional expectation:

$$f^\star_{\text{raw}} = \mathbb{E}[m_t ru + s\epsilon \mid \mathcal{I}] = m_t \mathbb{E}[ru \mid \mathcal{I}], \tag{5}$$

$$f^\star_{\text{unit}} = m_t \mathbb{E}[u \mid \mathcal{I}]. \tag{6}$$

□

Proposition 2. Let

$$T_{\text{raw}} = m_t ru + s\epsilon, \quad T_{\text{unit}} = m_t u + s\epsilon. \tag{7}$$

Gradients of squared error: $-2(T - f)$. Covariance of gradient:

$$\text{Cov}_{\text{raw}} = 4\,\text{Cov}(T_{\text{raw}}) = 4m_t^2 \text{Cov}(ru) + 4s^2 I, \tag{8}$$

$$\text{Cov}_{\text{unit}} = 4\,\text{Cov}(T_{\text{unit}}) = 4m_t^2 \text{Cov}(u) + 4s^2 I. \tag{9}$$

If r varies across samples, $\text{Cov}(ru) \gg \text{Cov}(u)$, increasing gradient variance.

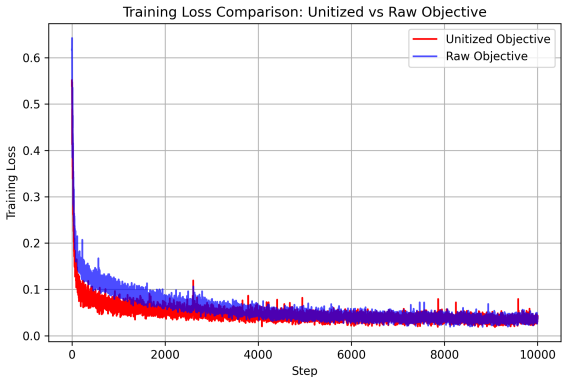

Figure 2. Training loss curves for raw and unitized objectives on SynthRAD2023. Unitized objective shows faster and smoother convergence on SynthRAD2023.

Sketch. Direct computation: for one sample $-2(T - f)$, variance of stochastic gradient is $\mathbb{E}[(T - \mathbb{E}T)(T - \mathbb{E}T)^\top] = \text{Cov}(T)$. Scaling by r inflates covariance proportionally to $\text{Var}(r)$. □

Corollary. For small training set n, sample variance of r can deviate from population variance:

$$\text{Var}_{\text{emp}}(r) \sim \mathcal{O}(1/n), \tag{10}$$

which amplifies stochastic gradient noise for raw objective. Unitization removes multiplicative scale, reducing gradient variance and improving stability.

Remarks. Standard SGD bounds: larger $\sigma^2 = \text{Var}(\text{grad})$ implies slower convergence. Raw objective has extra $\text{Var}(ru)$ term; small datasets exacerbate instability. □

C.2. Empirical Evaluation

To verify the theoretical predictions, we conduct experiments comparing raw and unitized objectives on two representative datasets: BraTS and SynthRAD2023. While the convergence difference on BraTS is modest, the benefit of the unitized objective is clearly observed on SynthRAD2023, where BBDM converges substantially faster.

The loss curves in Figure 2 further illustrate the effect: unitization leads to smoother and more stable convergence on SynthRAD2023.

Overall, we observed this effect somewhat incidentally: during training on SynthRAD2023, convergence was relatively difficult, and inspecting the loss curves revealed the advantage of the unitized objective. We hypothesize that the stronger benefit on SynthRAD2023 stems from its smaller dataset and higher variability between samples, which makes direct optimization of $y - x_0$ more challenging. In contrast, BraTS contains more abundant and consistent data, leading to stable convergence even with the raw

Method	SynthRAD2023				BraTS			
	NRMSE↓	PSNR↑	SSIM↑	I-SSIM↑	NRMSE↓	PSNR↑	SSIM↑	I-SSIM↑
MaskGAN	0.1526±0.008	17.6850±0.032	0.5883±0.027	0.9093±0.014	0.1035±0.003	20.1710±0.045	0.7028±0.016	0.9215±0.008
CT2MR	0.1138±0.002	19.2930±0.041	0.6015±0.015	0.9484±0.003	0.0643±0.003	23.8440±0.047	0.8535±0.007	0.9489±0.003
Dual.	0.1093±0.003	18.1780±0.037	0.6162±0.006	0.9382±0.004	0.0706±0.002	24.5100±0.042	0.8419±0.006	0.9404±0.004
I3Net	0.1653±0.014	11.7750±0.044	0.5581±0.038	0.9407±0.024	0.1502±0.010	18.9140±0.039	0.8074±0.019	0.9530±0.003
ALDM	0.1336±0.009	17.5610±0.046	0.6266±0.015	0.9368±0.013	0.0808±0.009	21.9210±0.043	0.7930±0.011	0.9544±0.003
mDAUNet	0.1069±0.012	18.6620±0.038	0.6331±0.008	0.9466±0.004	0.0602±0.004	24.6460±0.045	0.8378±0.009	**0.9597±0.003**
ReBrain	**0.1054±0.002**	**20.7590±0.045**	**0.6682±0.009**	**0.9662±0.003**	**0.0551±0.002**	**25.6070±0.042**	**0.8887±0.008**	0.9575±0.003

Table 1. Quantitative comparison under half-resolution input on SynthRAD2023 and BraTS datasets, with mean ± standard deviation over repeated runs.

Dataset	Raw Objective (h)	Unitized Objective (h)
BraTS	9.87	**9.42**
SynthRAD2023	8.15	**4.30**

Table 2. Training time comparison between raw and unitized objectives on BraTS and SynthRAD2023 datasets.

objective. In other words, when the sample set is small or heterogeneous, the scale of $y - x_0$ can vary considerably across examples, increasing gradient variance and slowing convergence. Unitization mitigates this effect by removing the multiplicative scale, resulting in more stable and faster training. For large and more homogeneous datasets like BraTS, the variability is naturally lower, so both objectives perform similarly. Overall, unitization proves effective for improving training stability and speed, especially in datasets with limited size or high sample variability.

D. More Results

Table 1 presents a quantitative comparison under half-resolution input on the SynthRAD2023 and BraTS datasets. Across all metrics, ReBrain consistently achieves state-of-the-art performance, outperforming existing methods such as MaskGAN, CT2MR, Dual, I3Net, ALDM, and mDAUNet. Most of the reported standard deviations are small, indicating that the results are stable with low fluctuation. These results highlight the effectiveness and reliability of ReBrain in producing high-fidelity reconstructions under sparse input conditions.

Besides the results shown in Figure 4, we additionally provide the comparison including ALDM and CT2MRI. The generated results are visualized in Figure 3, highlighting that our ReBrain method produces superior fidelity and structural consistency.

Impact of Similarity Between y and r. In our framework, the retrieved slice r is used to guide the generation conditioned on the input slice y. Typically, y and r are similar, enabling smooth and progressive synthesis. However, when the gap between y and r is large, the generation becomes highly stochastic, resulting in abrupt transitions

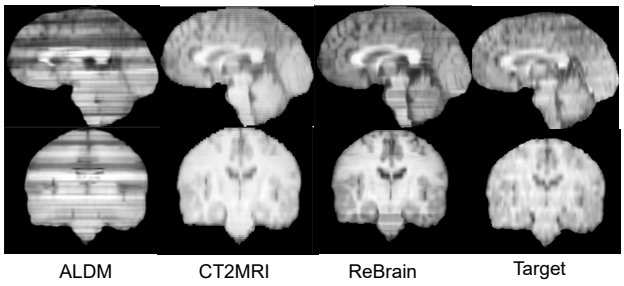

Figure 3. Comparison of generated results including ALDM, CT2MRI, and our ReBrain. ReBrain demonstrates superior fidelity and structural consistency.

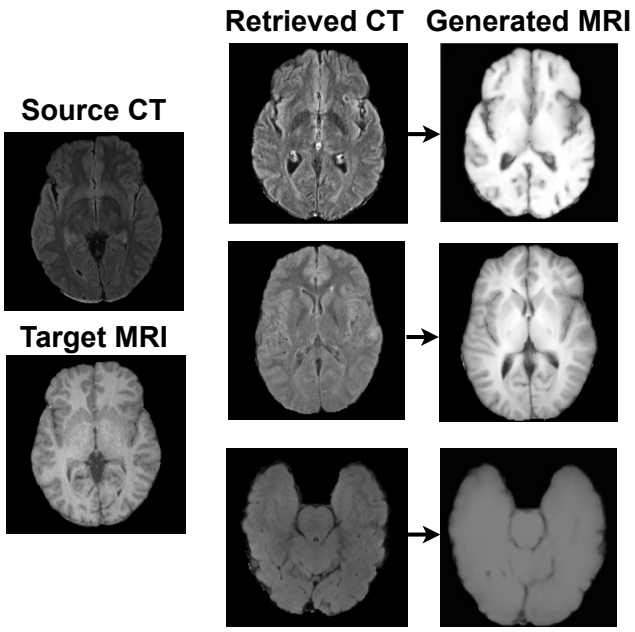

Figure 4. Illustration of uncertainty in generation caused by low similarity between input slice y and retrieved slice r

rather than gradual interpolation. As shown in Figure 4, we visualize multiple samples generated from a single y when paired with a poorly matched r. This highlights an interesting uncertainty phenomenon. Considering the deterministic

nature of medical imaging and the need to minimize risk, we employ interpolation instead of retrieval when the similarity between y and r is low.

Effect of Using Standard Diffusion with ControlNet. If we replace BBDM with a standard diffusion model while retaining ControlNet, the generated MRI tends to closely resemble r, with limited incorporation of y. This behavior is similar to directly feeding r into BBDM without ControlNet. When the discrepancy between the source slice y and the retrieved slice r is large, we consider the standard diffusion process starting from Gaussian noise a viable alternative. In such cases, sampling from noise yields more reliable reconstructions than initializing from y, which may impose incompatible semantic constraints on the generation.

E. Limitations

Technical considerations. We acknowledge that even when using y as the input to BBDM under the control of ControlNet, the generated structure still tends to approximate r. Although this approach improves upon directly using r as the input to BBDM, there remains room for further enhancement. This outcome does not fully align with our expectation of a combined effect of $y + r$. We speculate that this is related to the ControlNet training process, where the target MRI paired with inputs y and r is predominantly based on r. In future work, we will continue to optimize our method guided by these insights.

In some cases, the retrieved slice r has high similarity with the target slice y, but comes from a very different location in the brain. This causes a mismatch between appearance and true position, which may harm the quality of reconstruction. To reduce this problem, we added a simple check during retrieval: we remove slices whose positions differ too much from the query. This improves the trust in the results to some extent.

Ethical and practical considerations. Our framework uses previously observed, structurally similar slices only as a conditional reference to guide MRI reconstruction, rather than directly copying prior patient data. This ensures that retrieved slices act as a "soft hint," not a precise replication. While maintaining reasonable 3D structural continuity is one of the aims, it is achieved as part of balancing multiple reconstruction objectives rather than being the sole focus. SLERP interpolation is applied when retrieval fails or yields low-similarity slices, serving as a safe fallback that minimizes hallucinated structures.

Importantly, in clinical scenarios where the original CT lacks complete pathological information, no reconstruction method—including ours—can invent missing lesions. If a retrieved reference contains relevant pathology, it can assist in highlighting critical structures, potentially benefiting the patient. Conversely, if the patient is healthy or the reference has no pathology, any introduced errors are minor and limited to a careful review. Overall, this strategy substantially reduces hallucinations while keeping clinical risk very low, with potential benefits outweighing the unlikely costs.

To explicitly formalize the ethical trade-off between different types of errors, we define a weighted risk R as follows:

$$R = w_{\text{FN}} \cdot \mathbf{1}[y = 1 \wedge \hat{y} = 0] + w_{\text{FP}} \cdot \mathbf{1}[y = 0 \wedge \hat{y} = 1], \quad (11)$$

where $y \in \{0, 1\}$ denotes the true pathology label (1 = lesion, 0 = healthy), $\hat{y} \in \{0, 1\}$ the predicted label from reconstructed slices, w_{FN} is the weight for false negatives, and w_{FP} is the weight for false positives. We set $w_{\text{FN}} \gg w_{\text{FP}}$ to reflect that missing a lesion is clinically far more severe than mistakenly flagging a healthy region.

Without retrieval, sparse or continuous 3D reconstruction methods may produce high w_{FN} events, i.e., missing critical lesions. Our retrieval-augmented approach primarily shifts potential errors to w_{FP}-type events, which are less severe and can be mitigated through clinical review. This ensures that the most critical errors are minimized, while overall fidelity and structural consistency remain high, reducing potential harm to patients.

Moreover, we ensure that all reference slices used for guidance are fully anonymized, and no personally identifiable information is ever exposed. The retrieval mechanism is designed to prevent over-reliance on any single patient's data, further mitigating privacy risks. From an ethical standpoint, the model's predictions are intended as an assistive tool rather than a standalone diagnostic; final clinical decisions remain under the supervision of qualified professionals. Additionally, the model outputs are accompanied by uncertainty estimates and visual cues, promoting interpretability and enabling clinicians to verify reconstruction plausibility. These design choices collectively aim to respect patient privacy, reduce potential misuse, and maintain high standards of clinical responsibility.

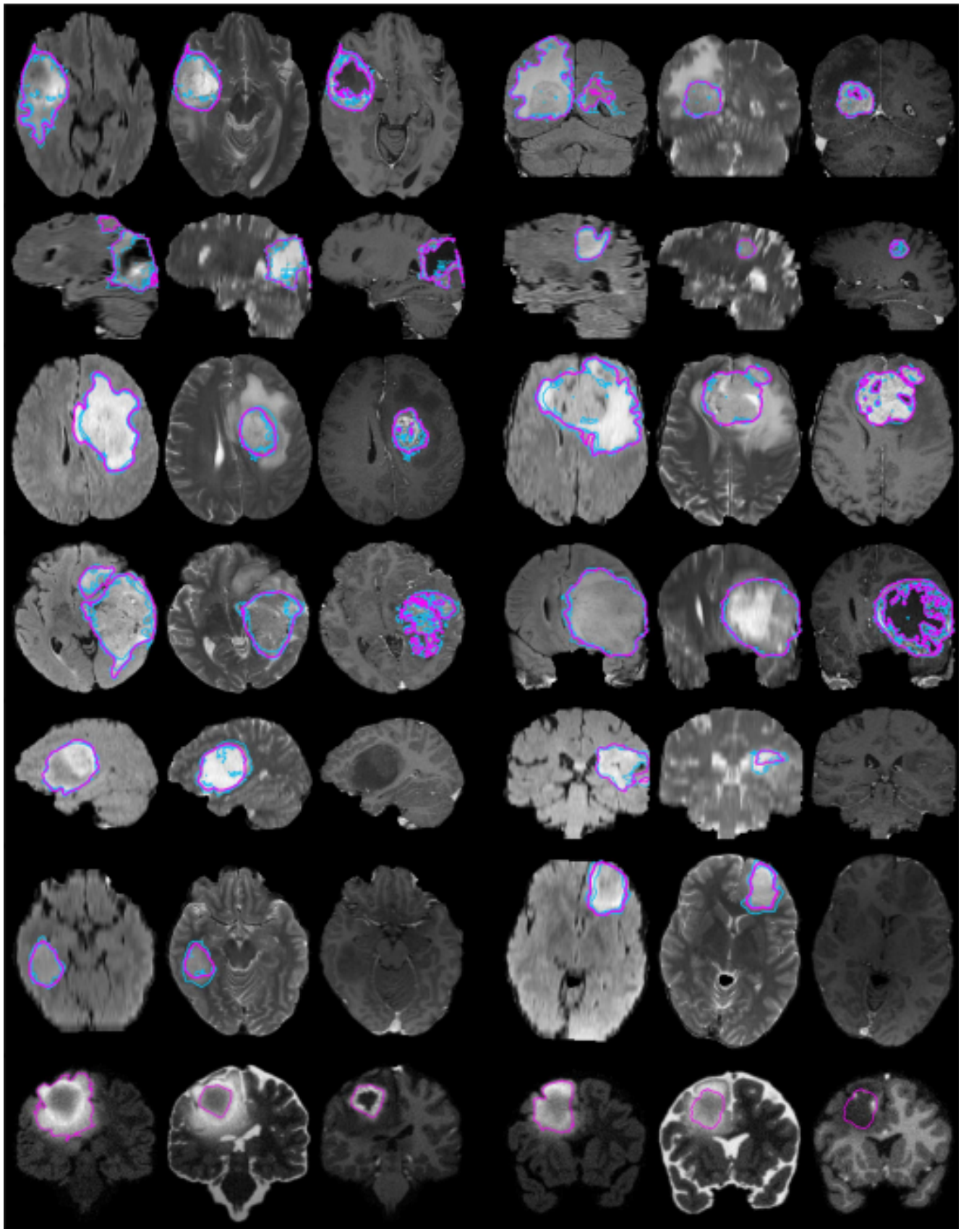

Figure 5. Examples from the BraTS training data, with tumor regions as inferred from the annotations of individual experts (blue lines) and consensus segmentation (magenta lines). Each row shows two cases of high-grade tumor (rows 1–4), low-grade tumor (rows 5–6), or synthetic cases (last row). Images vary between axial, sagittal, and transversal views, showing for each case: FLAIR with outlines of the whole tumor region (left); T2 with outlines of the core region (center); T1c with outlines of the active tumor region if present (right).

References

[1] Eungbean Lee, Somi Jeong, and Kwanghoon Sohn. Ebdm: Exemplar-guided image translation with brownian-bridge diffusion models. In *Computer Vision – ECCV 2024*, pages 306–323, Cham, 2025. Springer Nature Switzerland. 1

[2] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1952–1961, 2023. 1

[3] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. 1