Modality-Aware and Anatomical Vector-Quantized Autoencoding for Multimodal Brain MRI

A. Latent Utility

To assess how well different VAEs support downstream analysis, we compare sex-classification performance using either (i) the learned latent representations or (ii) the original MRI volumes. As summarized in Table 1, latent codes generally yield competitive or superior accuracy compared with image-based classifiers, despite operating at significantly lower dimensionality. NeuroQuant's latent embeddings consistently outperform other VAE baselines on both NCANDA and ABCD, confirming that its dual-stream design preserves discriminative anatomical information after compression. Although classification on original images remains slightly higher, as expected due to full-resolution inputs, the latent space offers clear advantages: accuracy variance across seeds is lower, and performance is more stable across modalities. This pattern indicates that NeuroQuant removes voxel-level noise and scanner-dependent variability, producing smoother and more reliable representations. Overall, these results demonstrate that NeuroQuant's latent space is not only reconstruction faithful but also highly informative and robust for downstream neuroimaging tasks.

B. Cross-modality Reconstruction Analysis

To further examine the anatomical fidelity of our cross-modal generation, Fig.1 visualizes SynthSeg segmentation overlays for two representative subjects across all three anatomical planes. The comparison between our cross-translated reconstructions and ground-truth volumes shows that NeuroQuant preserves region boundaries, cortical thickness patterns, and subcortical shapes across both T1→T2 and T2→T1 translations. In particular, major structures such as the cerebellum, ventricles, and deep gray nuclei remain well-aligned with their ground-truth counterparts, with minimal deformation or label leakage. This indicates that the anatomical stream successfully captures modality-invariant structure, allowing SynthSeg to segment generated images with reliability comparable to real scans. These results demonstrate that NeuroQuant maintains strong structural consistency even under modality translation, validating the quality of its disentangled anatomical representation.

Table 1. Quantitative comparison with state-of-the-art methods on sex classification. For each dataset and metric, the best latent representation result is highlighted in **bold**, the second-best result is underlined, and results by original images are in red.

		T1	T2
	Models	Accuracy	Accuracy
NCANDA	VQGAN	76.23	66.25
	SD-VAE	77.14	67.34
	MediTok	77.78	67.55
	MedVAE	78.12	67.87
	NeuroQuant (ours)	**79.09**	**68.36**
	- w/o joint training	78.57	67.95
	Original Image	81.73	70.19
ABCD	VQGAN	77.50	66.53
	SD-VAE	78.37	66.58
	MediTok	79.02	66.79
	MedVAE	79.31	67.13
	NeuroQuant (ours)	**80.34**	**67.58**
	- w/o joint training	79.79	67.18
	Original Image	80.88	68.34

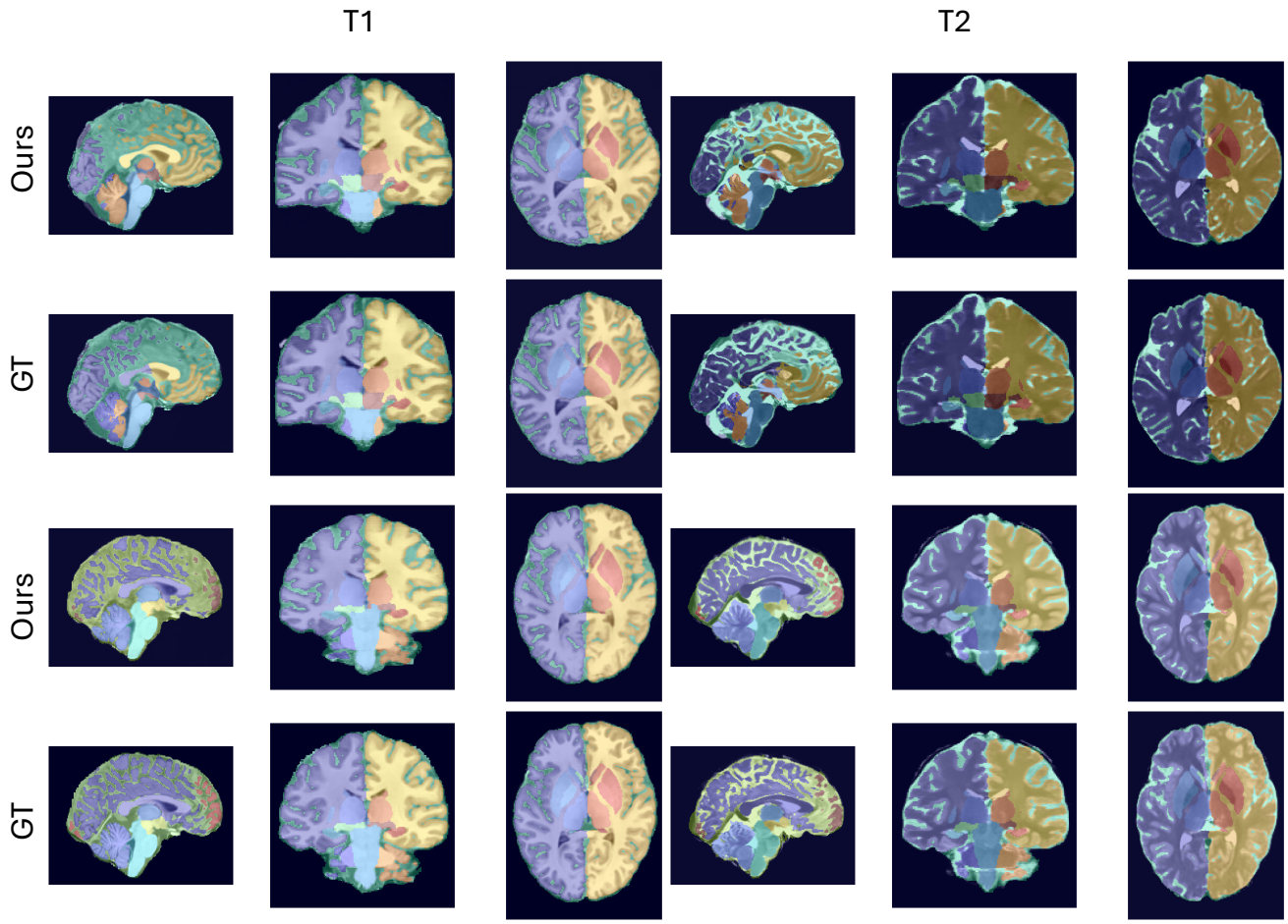

Figure 1. Cross-modal translation and SynthSeg visualization on two subjects. For each subject, an input T1 (or T2) volume is encoded into anatomical and modality-specific latents, and the decoder reconstructs the target modality (T2 or T1). SynthSeg segmentation maps (half-overlay) are shown for three orthogonal views (sagittal, coronal, axial). "GT" refers to SynthSeg applied to the ground-truth volume; "Ours" indicates SynthSeg applied to NeuroQuant's cross-translated output.