Supplementary Materials for Controllable Text-to-Image Synthesis for Multi-Modality MR Images

In this supplementary document, we provide a comprehensive extension to the primary paper by detailing the models employed in our experiments and presenting further qualitative results. Section 1 delineates the hyper parameters of the implemented models. Subsequently, Section 2 exhibits additional synthetic MR images accompanied by corresponding cross-attention maps. Section 3 describes the details of the Turing Test used for qualitative evaluation.

1. Model Implementation and Training

The model was trained using PyTorch with an NVIDIA A100 GPU. The rest of the parameters are available at Table 1.

Parameter	Value
Image size	256
Learning rate	1e-6
Batch	8
Diffusion steps	100
Channels	64
Heads	4
Heads channels	8
Attention resolution	4,2,1
Num Resblocks	2
U-Net Image size	64
Prompt dimension	768
CLIP model	ViT-L/14

Table 1.

2. Additional visualization of Synthesis Results

This section provides additional visualization of synthetic images and the corresponding cross attention maps $H_k[h, w]$ (h=256, w=256) to individual words on the prompt by suggested framework. Figures 1, and 2 demonstrate the results of the synthetic MR images featuring IDH mutant and wild generated via a guided approach utilizing both input prompts and structural masks.

“A {*} modality MR image, type of mutant glioma”

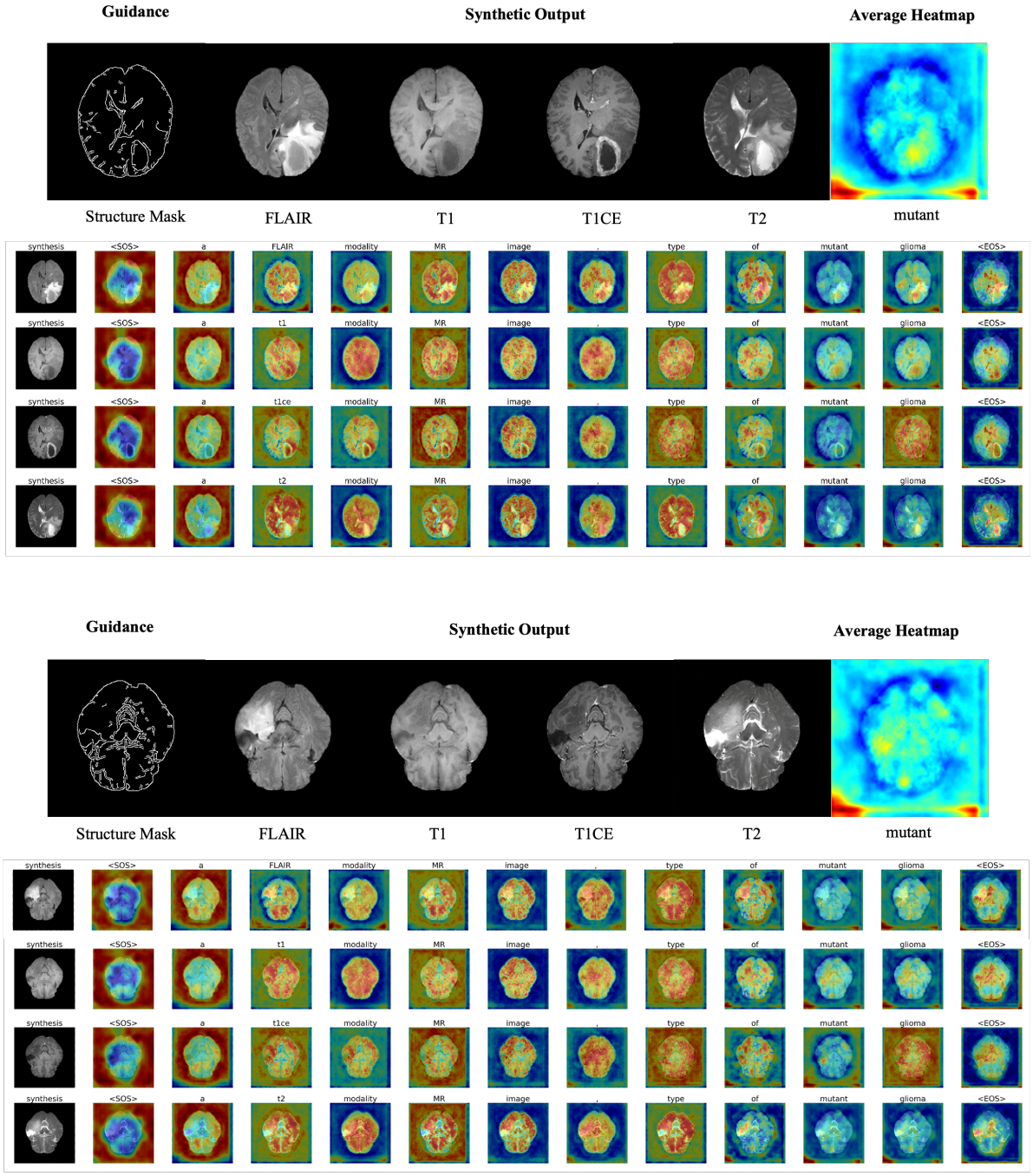
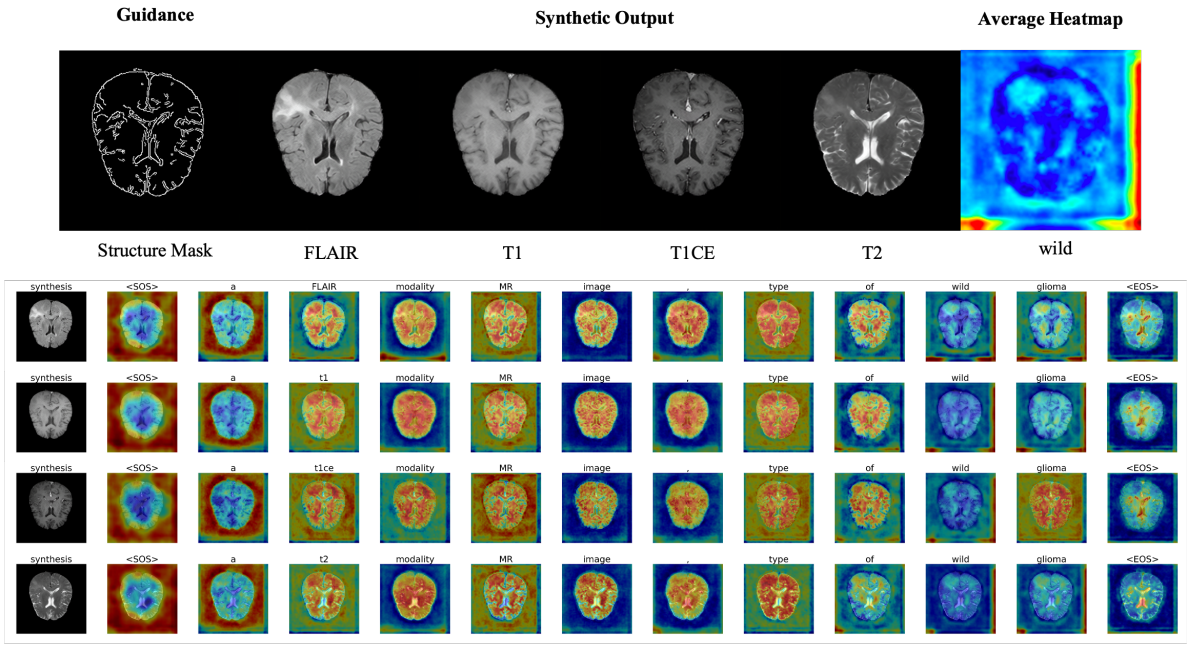


Figure 1. Synthetic MR images featuring IDH mutant generated via a guided approach utilizing both input prompts and structural masks. The input prompt uses an asterisk {*} as a placeholder to accommodate various MR modalities, such as FLAIR, T1, T1CE, and T2. This process is complemented by the visualization of attention maps to understand the model’s focus during image synthesis.

“A {∗} modality MR image, type of wild glioma”



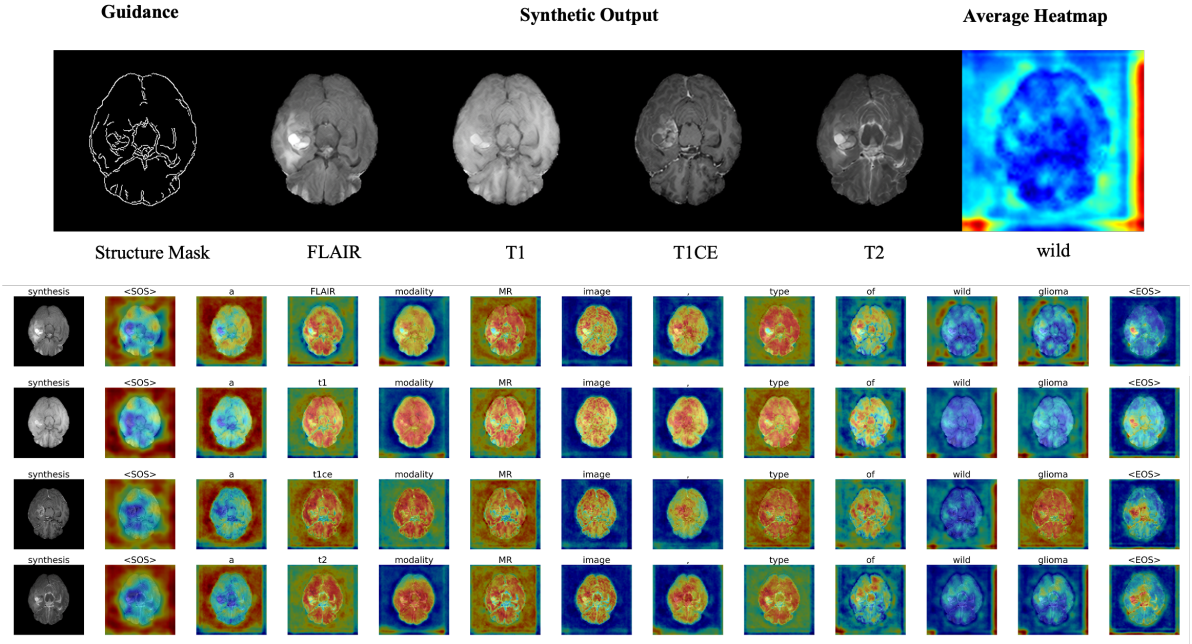


Figure 2. Synthetic MR images featuring IDH wild generated via a guided approach utilizing both input prompts and structural masks. The input prompt uses an asterisk {*} as a placeholder to accommodate various MR modalities, such as FLAIR, T1, T1CE, and T2. This process is complemented by the visualization of attention maps to understand the model’s focus during image synthesis.

3. Details of Qualitative Evaluation

The Turing Test for qualitative evaluation is structured into two distinct type. Type 1 involves a comparative analysis in which experts are presented with a random pair of real and synthetic MR images from the same modality and are tasked with identifying the synthetic one. Type 2 entails a qualitative assessment where experts rate a sequence of generated MR images on a 0 to 5 scale. In this second type, evaluators also consider whether the sequence accurately represents the appearance of a tumor with an IDH mutant-like or wild-like phenotype. Each expert assesses 20 items per category, with the evaluation capped at 30 minutes. Table 2 details the specific results for each category of the test. Our test and image samples are available online. Type 1, comparative analysis at quilgo.com/t/krcysBoBRO99C7Kc, and Type 2, the qualitative assessment can be found at quilgo.com/t/y4SpjnCUSKcJhpLh.

Reviewer	Average results per test type		
	Real vs. Synthesis accuracy	Image Scoring	Mutant vs. Wild accuracy
#1	40%	4.05 ± 0.6863	65%
#2	55%	4.7 ± 0.6569	70%
Average	**47.5%**	$\mathbf{4.375 \pm 0.7403}$	**67.5%**

Table 2. Average scores segmented by test type and reviewer

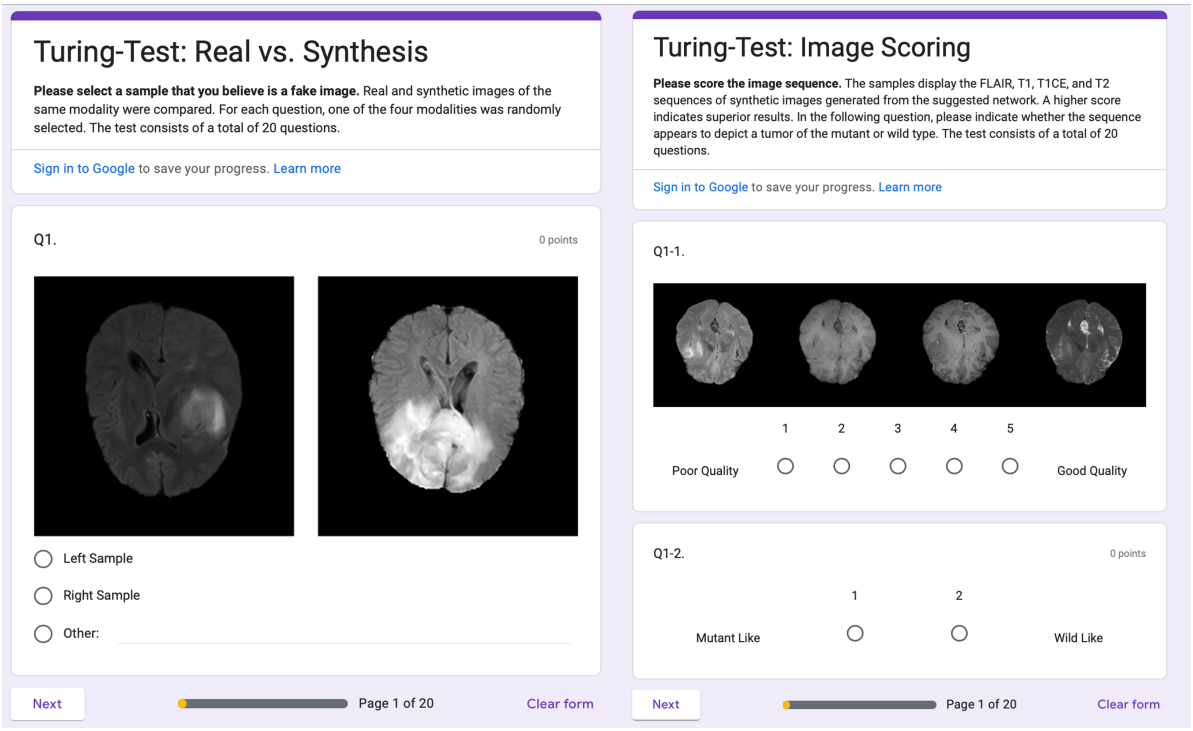


Figure 3. Illustration of our Human evaluation Turing-test created for clinical experts to evaluate synthetic images. (Left: Type 1, Right: Type 2)