A Additional Training and Architectural Details

Our model (segmentation-guided diffusion). The denoising model (UNet)'s encoder is constructed from six standard ResNet down-sampling blocks, with the fifth block also having spatial self-attention, with $(128, 128, 256, 256, 512, 512)$ output channels, respectively. The decoder is simply the up-sampling reverse of the encoder. We use a standard forward process variance schedule that linearly increases from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ [?]. For training, we use the AdamW optimizer [?] and a cosine learning rate scheduler [?] with an initial learning rate of 10^{-4}, with 500 linear warm-up steps. We train for 400 epochs with a batch size of 64 (about 26 hours), and we perform all training and evaluation on four 48 GB NVIDIA A6000 GPUs. We use the Diffusers library as a backbone (`https://github.com/huggingface/diffusers`).

A.1 Comparison models

SPADE. We train SPADE [?] using the default settings, with a batch size of 128 for 50 epochs. We did not adopt the changes of the recent brain MRI SPADE model [?] because they are not applicable to our datasets/task, namely: (1) the contrast-based clustering is not applicable due to us using pre-contrast MRIs or CT, (2) we work with standard categorical segmentation maps, not partial volume/probabilistic segmentation maps, so changes using the latter are not applicable, and (3) we work with independent 2D slice images, rather than full 3D volumes, so the enforcement of style and content separation via using different slices from the same volume during training is not applicable.

ControlNet. We adapted ControlNet [?] to each of our medical image datasets as was instructed at their official tutorial (`https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md#sd_locked`) for use with datasets that are out-of-distribution (*e.g.*, medical images) from their model's very large natural image pre-training set, using empty prompts for text inputs. We note that despite this tutorial, none of this was tested in the ControlNet paper, which may explain ControlNet's poor performance on our medical datasets.

This involved first finetuning the VAE for 200 epochs, then finetuning the Stable Diffusion (SD) model for 400 epochs using the respective breast MRI or CT organ training set images. We then finetuned the ControlNet with the images and their corresponding masks for segmentation guidance for 200 epochs. The pretrained (pre-finetuning) models are from the SD v1.5 checkpoints available on Hugging Face at `https://huggingface.co/runwayml/stable-diffusion-v1-5`. For all training, we set the batch size to 128, the initial learning rate to 10^{-4}, and adopted cosine annealing learning rate schedulers rate with 500 steps of warm-up.

A.2 Auxiliary segmentation model

We used the MONAI UNet (`https://docs.monai.io/en/stable/networks.html`) with 1-channel input and (number of target object classes + 1)-channel output. The sequence of intermediate UNet channels was set to (16, 32, 64, 128, 256). We trained each model for 100 epochs with a batch size of 8 and selected the models with the lowest validation loss, with an initial learning rate of 10^{-3} and a cosine annealing scheduler.

B Additional Sampled Images

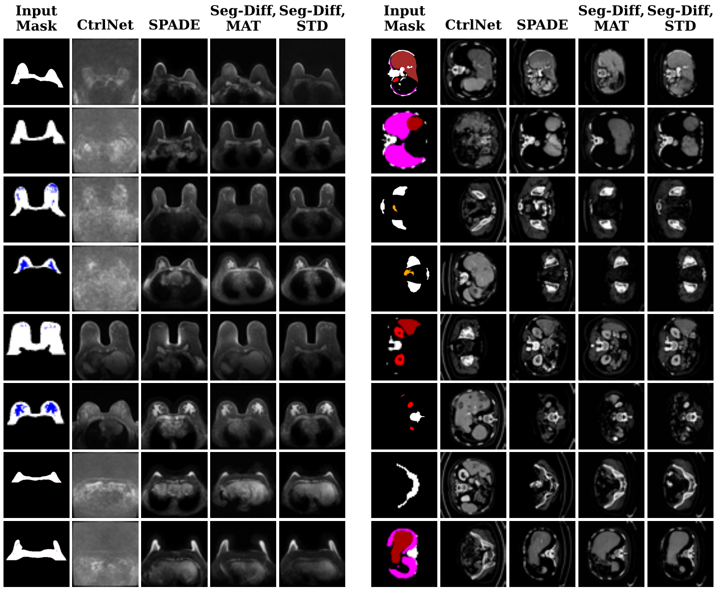

Fig. 5. Additional samples from all segmentation-conditional models; breast MRI on the left, CT organ on the right. Please see Fig. 2 caption for more details.

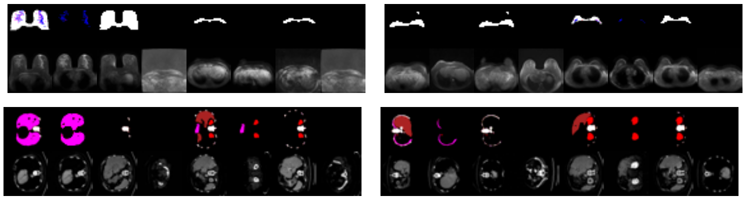

Fig. 6. Additional samples from our mask-ablated-trained model with various classes removed from given input segmentations for breast MRI (top) and CT Organ (bottom).