**Supplemental Material** In this section, we provide more experiment results and visualization. The results will be presented according to the experiments.

## 1 Supplementary results for FID and Dist-FID metrics

The following three charts in Figure 1 depict the best epoch model selection strategy. The Dist-FID shows the consistent best epoch score compare to the original FID. By contrast, the rest selection strategy could not accurately reflect the optimized model. FID_1% and FID_5% scores show the FID using the given percentage subset of all data. The FID of each source shows the scores calculated from the given local dataset only. The black arrows indicate the optimized epoch for FID and Dist-FID. The FID is the ideal scores to identify the optimized epoch generative model with the best synthetic image quality. However, simply applying the FID inception feature calculation to the distributed private datasets tend to be impractical since the data privacy restriction and significant bandwidth of inception feature accumulation for all real-synthetic image pairs. It's also worth to notice that the more heterogeneity there is in the cohort, the greater fluctuation between different medical image centers. No matter the heterogeneity of each experiment, Dist-FID is always consistent with the FID score which calculated from all data together.

## 2 Supplementary results on each experiments with the box plots

Figure 2 shows the DICE, HD95, SD and ACC scores across four experiments. DSL demonstrates superior performance compare to other model trained with local data and other distributed GAN methods.

## 3 Supplementary results of the variation of synthetic images by DSL

We show some examples by using such test-time augmentation (TTA) in the supplementary figures 3, 4 and 5. DSL's implementation provides random noise in the form of dropout, applied on several layers of generator at both training and testing. Therefore, there is minor stochasticity in the synthetic images when repeating the same input. However, we can make DSL generate more diverse images by applying random transformations to the input in testing, such as random scaling, shift, flip, and so on. We can see that by applying random TTA on the input label images, the generator produces varied contexts in many detailed areas. Due to the random transformations of the input, the synthetic images may have slightly different scales and positions (flip if any was reversed for better visualization).

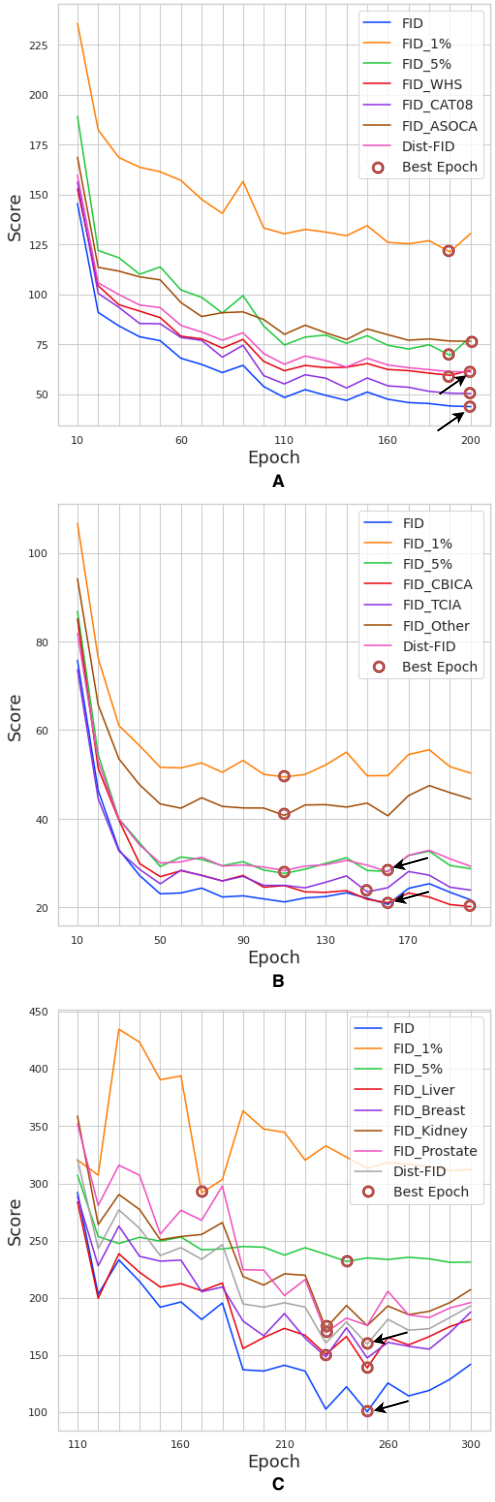


Figure 1: The complete set of FID and Dist-FID scores in three experiments. (A): Results for the DSL with the cardiac dataset. (B): Results for the multi-modality DSL with the Brain BraTS dataset. (C): Results for the continuous-learning DSL with the Nuclei dataset.

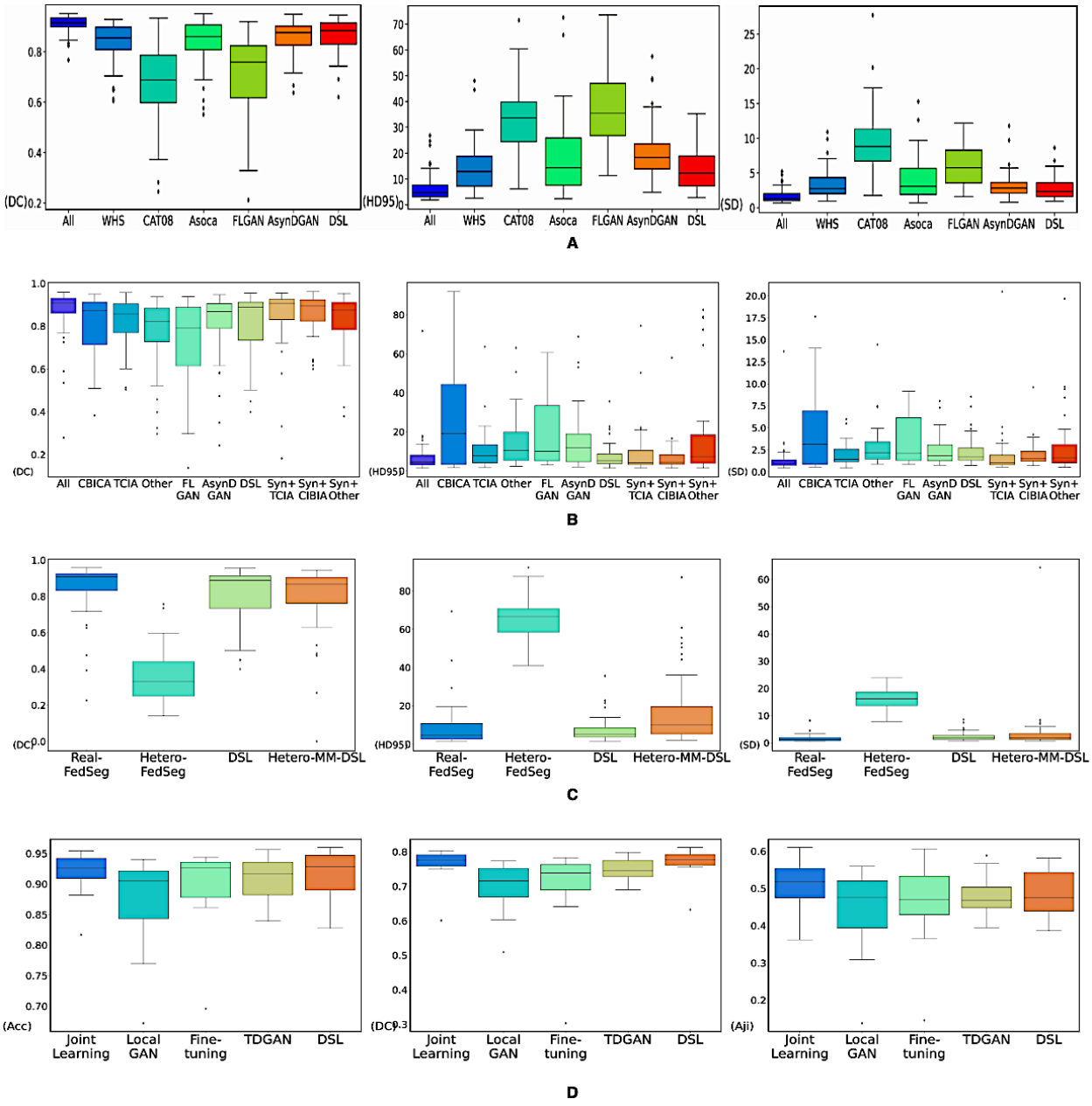


Figure 2: The details of each experiments with the box plots. (A) shows the result of DSL with cardiac CTA images. (B) shows the result of multi-modality DSL with brain BraTS dataset. (C) shows the missing modality completion DSL's result. (D) shows the continuous learning DSL's result.

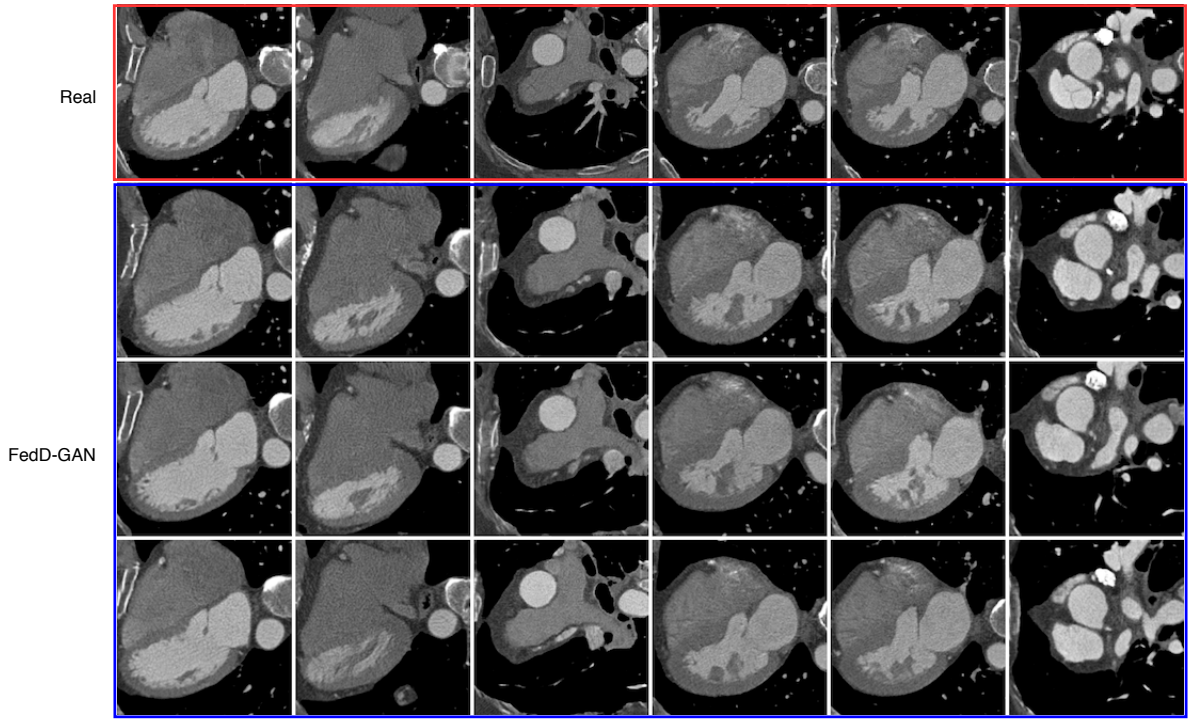


Figure 3: The first line shows six real cardiac CT images, and the following three rows show the corresponding synthetic images generated by DSL for three times.

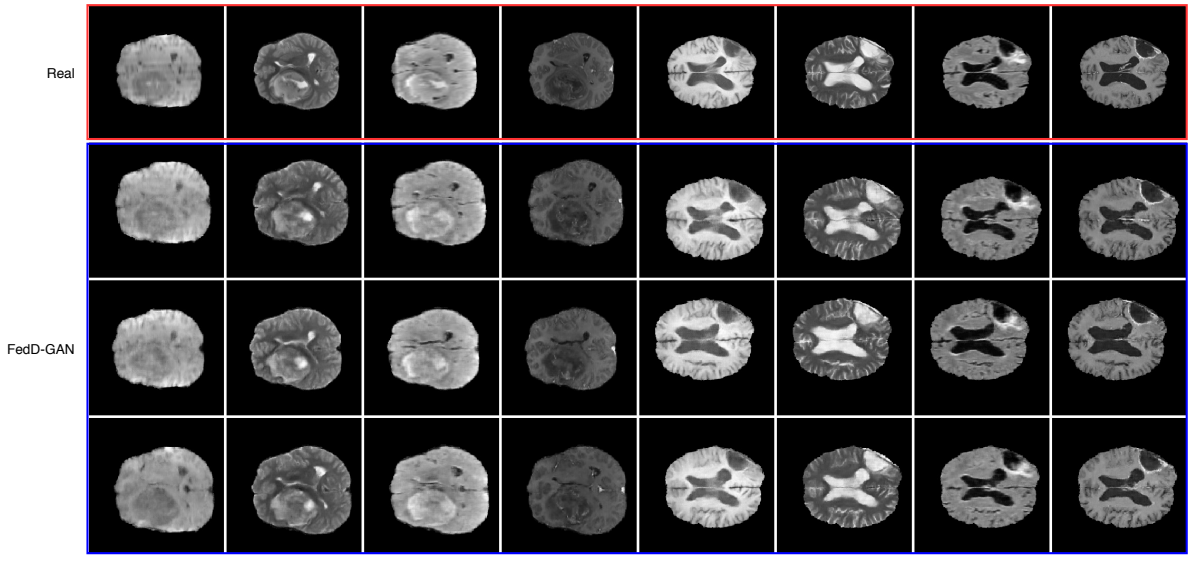


Figure 4: The first line shows the real brain MR images (including four modalities T1, T2, Flair, T1c) of two cases, and the following three rows show the corresponding synthetic images generated by DSL for three times.

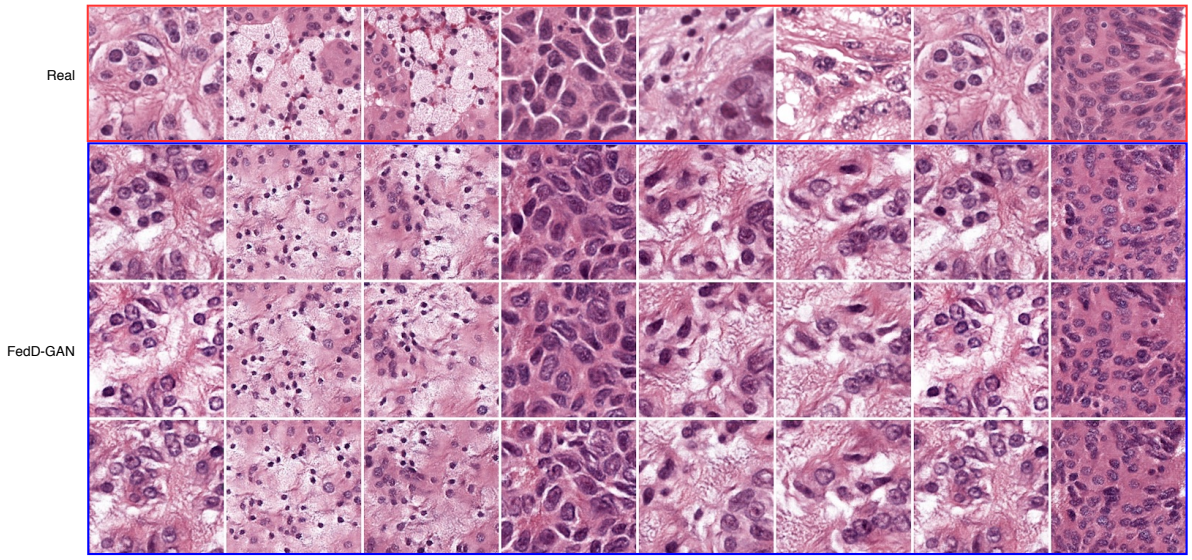


Figure 5: The first line shows eight real histopathology image patches, and the following three rows show the corresponding synthetic images generated by DSL for three times.