1 Responce of ICML submission 3891

	Generator 1	Generator 2	Generator 3	Generator 4
FID-avg	Superior (270.54)	Superior (280.38)	Superior (268.69)	Inferior (310.69)
KID-avg	Inferior (0.467)	Inferior (0.480)	Inferior (0.461)	Superior (0.261)

Figure 1: The counterintuitive ranking given by FID-avg score on FFHQ dataset. As we discussed in Fig. 6 in main text, FID-avg even gives better scores to collapse models as we shown in generators 1 to 3, where we set truncation $\tau = 0.1$ as we explained in Section 5.3.

	Generator 1	Generator 2	Generator 3	Generator 4
FID-avg	Superior (150.22)	Superior (156.46)	Superior (147.50)	Inferior (175.60)
KID-avg	Inferior (0.217)	Inferior (0.228)	Inferior (0.212)	Superior (0.118)

Figure 2: The counterintuitive ranking given by FID-avg score on AFHQ dataset.

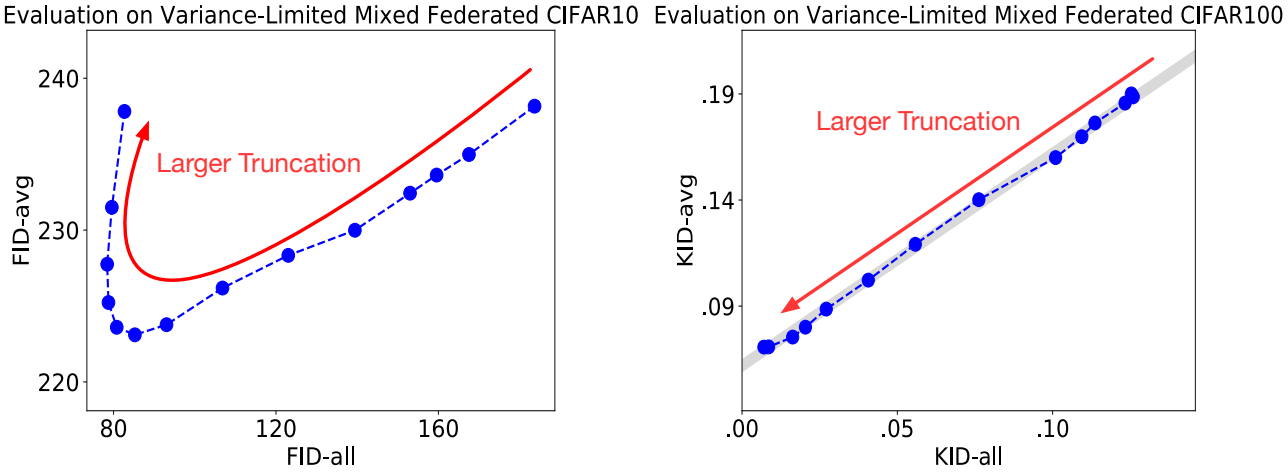

Figure 3: Evaluation results on CIFAR10 where clients hold samples from two classes. The setting is similar to that in Fig. 16 in main text but with sample overlap.