On the Distributed Evaluation of Generative Models

Supplementary Material

7. General 1-Wasserstein-Distance evaluation Metrics

Let $\mathbb{P}_g$ and $\mathbb{P}_t$ represent the distribution of generated set and training set. The Wasserstain-1 distance between $\mathbb{P}_g$ and $\mathbb{P}_t$ is,

$$W(\mathbb{P}_g, \mathbb{P}_t) = \inf_{\lambda \in \Pi(\mathbb{P}_g, \mathbb{P}_t)} \mathbb{E}_{(x,y)\sim\lambda}[\|x - y\|], \quad (4)$$

where $\Pi(\mathbb{P}_g, \mathbb{P}_t)$ denotes the set of all joint distribution $\lambda(x, y)$ whose marginal distribution are respectively $\mathbb{P}_g$ and $\mathbb{P}_t$. However, the direct estimation of $W(\mathbb{P}_g, \mathbb{P}_t)$ is highly intractable. On the other hand, the Kantorovich-Rubinstein duality [36] gives,

$$W(\mathbb{P}_g, \mathbb{P}_t) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_g}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_t}[f(x)], \quad (5)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{R}^n \to \mathcal{R}$. Therefore, if we have a parameterized family of functions $\{f_\theta\}_{\theta\in\Theta}$ that a 1-Lipschitz, we could considering solve this problem,

$$\max_{\theta\in\Theta} \mathbb{E}_{x\sim\mathbb{P}_g}[f_\theta(x)] - \mathbb{E}_{x\sim\mathbb{P}_t}[f_\theta(x)]. \quad (6)$$

To estimate the supremum of Equation (5), we employ a family of non-linear neural network f_θ which are repeatedly stacked by the fully connected layer, the spectral normalization and RELU activation layer. There are three repeated blocks in the network f_θ and the last block does have RELU. The feature is extracted by pre-trained Inception-V3 network. By optimizing the parameters in f_θ to maximize $\mathbb{E}_{x\sim\mathbb{P}_g}[f_\theta(x)] - \mathbb{E}_{x\sim\mathbb{P}_t}[f_\theta(x)]$ over $\mathbb{P}_g$ and $\mathbb{P}_t$, we can finally get an estimation of $W(\mathbb{P}_g, \mathbb{P}_t)$. And similarly, we can also define average score W-avg and collective-data-based score W-all under the distributed learning setting. Similar to the CIFAR100 experiment in the main body of paper, we extracted samples from each single class of CIFAR100 and evaluate these samples on federated CIFAR10 dataset. We illustrate a subset of W-avg / W-all pairs in Figure 7. According to experiment results, we find that general 1-Wasserstein-Distance evaluation metric also shows inconsistent behaviours in the distributed evaluation settings.

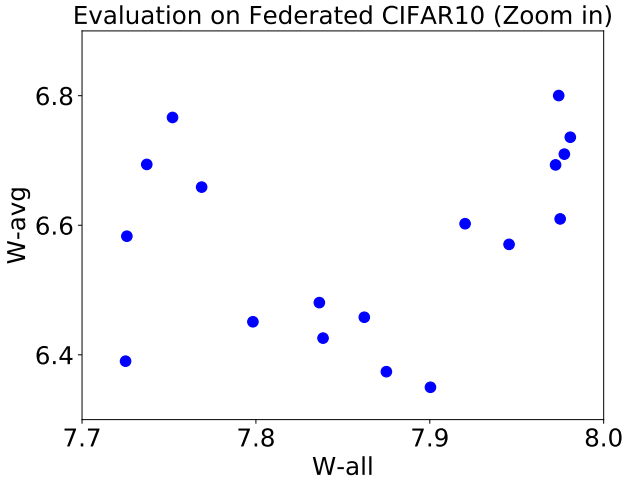


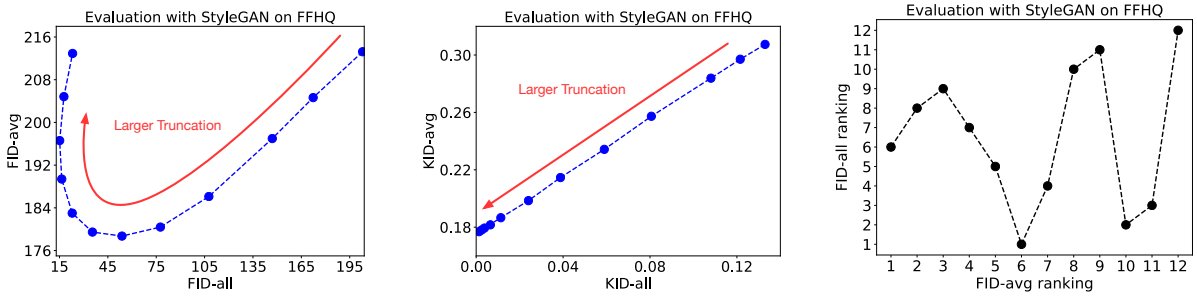
Figure 7. Evaluation with 1-Wasserstein-Distance on Federated CIFAR10.

8. Extra Experiment Results on FFHQ and AFHQ

8.1. Complex FFHQ Setting

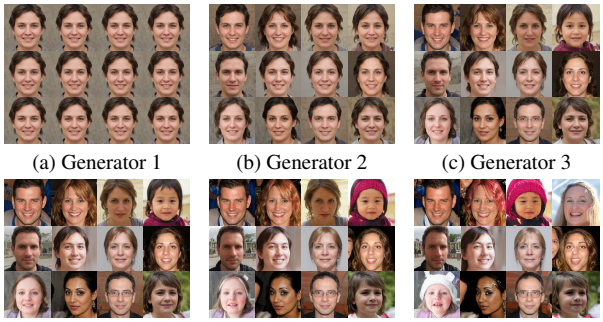
Experiment Setting. Following a similar methodology as described in the previous subsection, we synthesized a series of diversity-controlled generators by applying the standard truncation technique [18] to the random noise vector $\mathbf{z}$. We varied the truncation factor τ over $[0.01, 1.0]$. The effect of changing the truncation factor on the generated samples' diversity is illustrated in Figure 9. For every attempted τ, we generated 50K samples. Additionally, to simulate a distributed setting with heterogeneous data distributions, we simulated 100 clients, each with images synthesized with truncation factor $\tau = 0.25$. The centers of image distributions for each client varied, resulting in intra-client similarity and inter-client diversity as depicted in Figure 10. We evaluated the generators in the distributed setting of the simulated clients using both the discussed FD and KD-based evaluation scores.

Numerical Results. As shown in Figure 8(a) and Figure 8(b), FD-avg and KD-avg led to different rankings of the models. The plot of FD-avg versus FD-all led to a U-shaped curve, revealing inconsistent rankings of the models. On the other hand, the difference between KD-avg and KD-all remained constant for the generators, as shown in Theorem 1. These findings are also visible in the comparative rankings based on FD scores in Figure 8(c). The results



(a) FD-scores (b) KD-scores (c) Rankings given by FD-scores

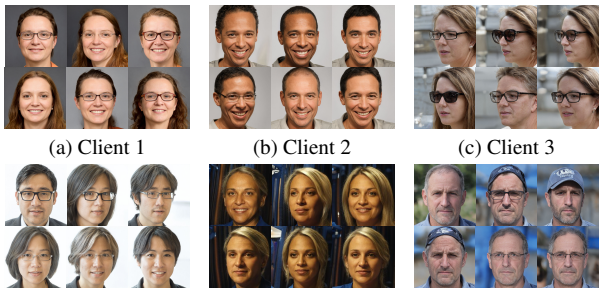
Figure 8. The results of evaluating generators in Figure 9 over clients in Figure 10.



(a) Generator 1 (b) Generator 2 (c) Generator 3

(d) Generator 4 (e) Generator 5 (f) Generator 6

Figure 9. Illustration of generators with different truncation factors. From (a) to (e), $\tau = 0.01, 0.2, 0.4, 0.6, 0.8, 1.0$, where τ is the truncation parameter.



(a) Client 1 (b) Client 2 (c) Client 3

(d) Client 4 (e) Client 5 (f) Client 6

Figure 10. Illustration of simulated clients with heterogeneous distributions via truncation technique.

suggest that although KD scores and FD-all prefer generators with higher truncation factors and diversity, FD-avg preferred generated data with limited diversity matching the bounded diversity level at each client.

8.2. Results on AFHQ dataset

Utilizing a another model weight that is pre-trained on AFHQ-wild dataset [4], we extend the above experiment. We gradually increase the truncation parameter of generators We show some examples from generators, clients on Figure 12 and Figure 11 and plot the relationship between FD-scores and KD-scores on Figure 13. The results that are illustrated in the figures still indicates that FD-avg can lead to different ranking with FD-all, while KD-avg can be treated as a more stable evaluation metric in the distributed learning setting.

8.3. Utilized Clients in Distributed Optimization via MMD^2

In Section 5.5, we fine-tune pre-trained StyleGAN models across multiple clients using the MMD^2 distance in a distributed manner. The clients we employ are generated from StyleGAN2, pre-trained on the FFHQ dataset, using the truncation method as introduced in the main text. This subsection provides further details and sample images from the utilized clients.

Training Details. We follow the standard training protocols of StyleGAN2-ADA, incorporating two primary loss functions. The first is the standard GAN training loss, applied to both the generated samples and the original FFHQ dataset. The second is the MMD^2 distance, computed using a polynomial kernel of order 3, applied between the generated samples and client samples. To balance these losses, we set the MMD^2 distance weight factor to 5. Training is conducted for 1,000k images. Additional hyperparameters, including learning rate and data augmentation settings, follow the "paper256" configuration from the official implementation.

Client Samples. We randomly selected several samples from each client, as shown in Figure 14. In total, we use 12 clients for each experiment (glasses and head accessories).

9. Evaluation on Synthetic Gaussian Mixture

As discussed in main text, the optimal selection of the covariance matrix differs for the FD-all and FD-avg aggregate scores. To illustrate this distinction, we performed a toy experiment, revealing that FD-avg attains its minimum value when the generator's variance closely approximates that of

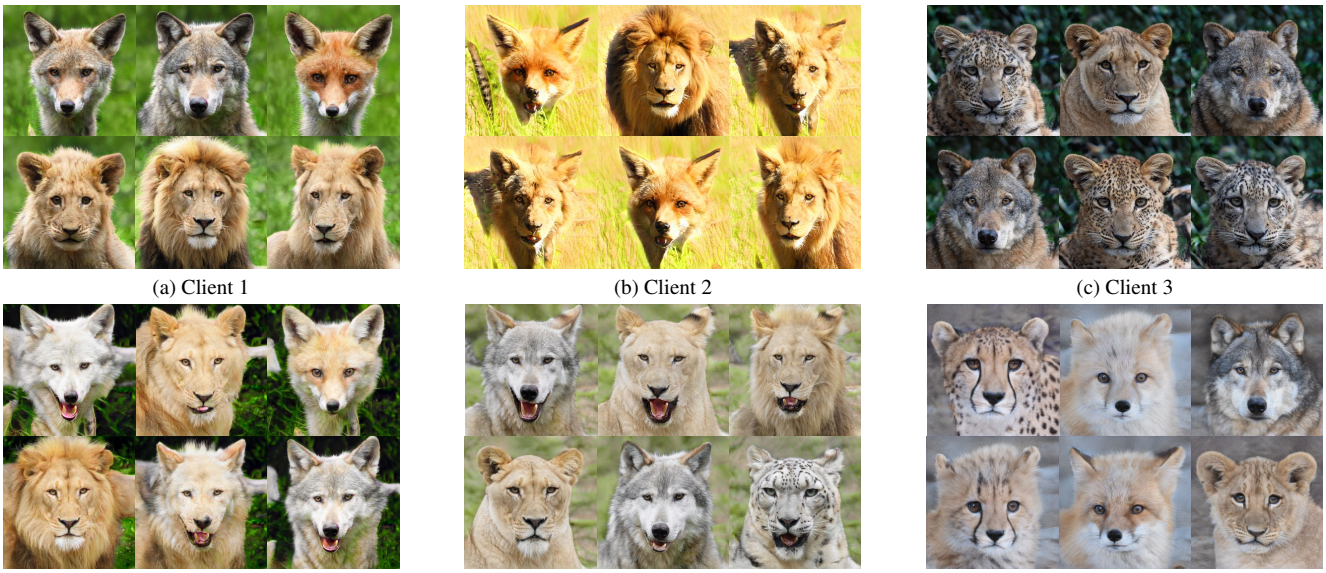
Figure 11. Illustration of randomly-generated samples from the variance-controlled generators on AFHQ dataset. Images in each row are synthesized with the same truncation parameter.

an individual client, whereas FD-all is minimized when the variance equals that of the aggregate distribution.

Setup. Our experimental setup involves two clients, denoted by C_1 and C_2. C_1 possesses a dataset consisting of 50,000 samples drawn from the Gaussian distribution $\mathcal{N}([1,0]^\top, \mathbf{\Sigma})$, while C_2 holds a dataset with 50,000 samples drawn from $\mathcal{N}([-1,0]^\top, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \text{diag}([1,1]^T)$. We introduce a generator, denoted as G_{var_x}, which is parameterized by var_x. var_x regulates the variance of the generator along the X-axis. Specifically, G_{var_x} generates 50,000 data points following a Gaussian distribution $\mathcal{N}([0,0]^T, \mathbf{\Sigma_G})$, where $\mathbf{\Sigma_G} = \text{diag}([\text{var}_x, 1]^T)$. The relationship between the two clients and the generator is visually depicted in Figure 15. Additionally, we introduce an "ideal estimator" denoted as $\hat{E} = C_1 \cup C_2$. This ideal estimator possesses the unique ability to replicate the distribution of the training dataset perfectly. We employ the ideal estimator as a reference for our analysis.

Evaluation Metrics. We measure the similarity between samples generated by clients and generators using the Fréchet distance (FD), which follows from the Wasserstein-based definition of FD-all and FD-avg without the application of the pre-trained Inception network. We consider the aggregate scores FD-avg and FD-all as defined in Equation (2) and Equation (1). Note that the FD-all for the ideal estimator is zero and we use FD-ref $= \frac{1}{2}\sum_{i=1}^{2} \text{FD}(\hat{E}, C_i)$ as a reference for FD-avg. We also measure the Kernel distance (KD), which follows the definition of KD-all and KD-avg without Inception network. KD-ref is defined for the kernel distance in a similar fashion to FD-ref.

Results. By increasing var_x from 0 to 4, we get a sequence of FD-avg / FD-all pairs and we plot them with the var_x in Figure 15. Our experimental results highlight the following conclusions. First, we observed that the minimum of FD-all occurs at $\text{var}_x = 2$, while that of FD-avg occurs at $\text{var}_x = 1$, which indicates that the optimal solutions of var_x to minimize FD-all and FD-avg are inconsistent. In this case, FD-all and FD-avg lead to different rankings of the models with $\text{var}_x = 1$ and $\text{var}_x = 2$. Addition-



(a) Client 1 (b) Client 2 (c) Client 3

(d) Client 4 (e) Client 5 (f) Client 6

Figure 12. Illustration of random samples from randomly selected variance-limited clients on AFHQ dataset.

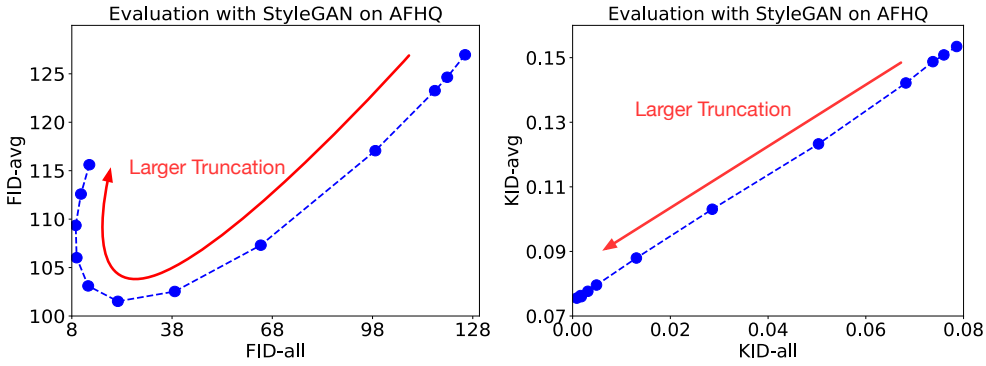


Figure 13. Evaluation on AFHQ dataset with StyleGAN2.

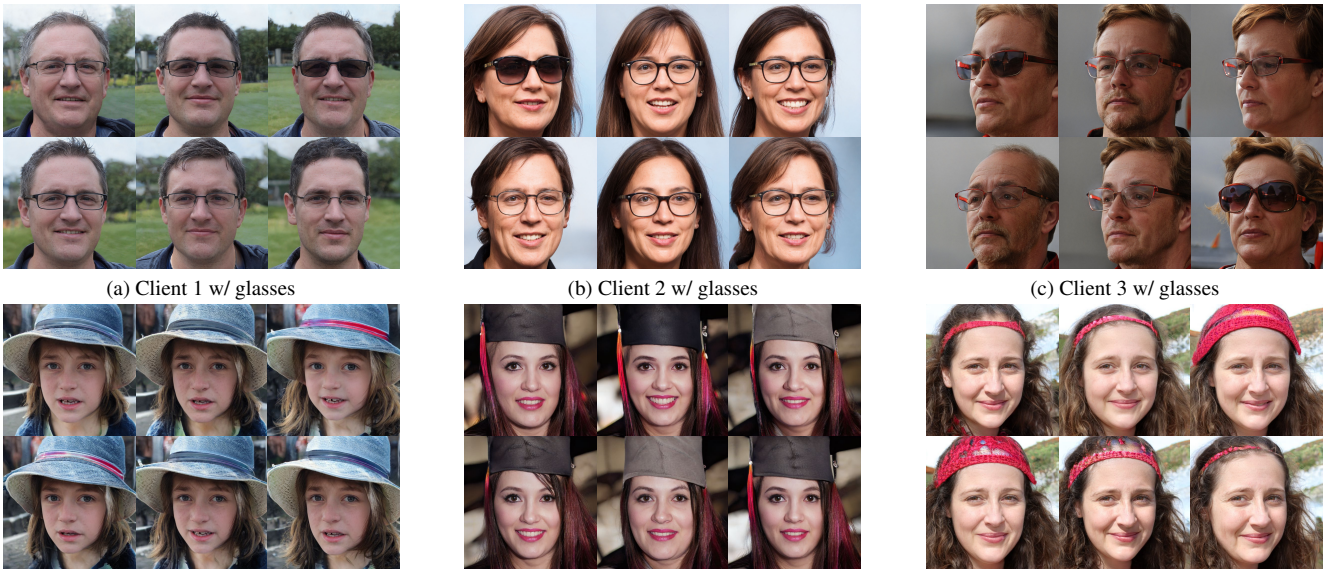
ally, we observed that, counterintuitively, the 'ideal estimator' did not reach the minimum average of the Fréchet distances. The distance between KD-avg and KD-all remains the same with the change of var_x and both of which reach minimum at $\mathrm{var}_x = 2$. The toy experiment highlights how a co-variance mismatch between clients and the collective dataset leads to inconsistent rankings according to aggregate Fréchet distances.

The Log-Likelihood Score. We also evaluated the synthetic Gaussian mixture dataset with the standard log-likelihood (LL) score. In this experiment, we note that we have access to the probability density functions (PDF) of the simulated generator. We utilized the generator G_{var_x} described in the main text and performed the evaluation over the parameter var_x in the range $[0, 40]$. As can be shown in the general case, LL-avg and LL-all led to the same value for every evaluated model. As shown in Figure 16(a), they reached their maximum value at $\mathrm{var}_x = 2$. On the other hand, we set a new generator G_{mean_x} generating samples according to $\mathcal{N}([\mathrm{mean}_x, 0]^\top, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathrm{diag}([2, 1]^T)$. We gradually increased mean_x from -2 to 2 and plotted LL-avg, LL-all, and LL-ref in Figure 16(b).

10. Evaluation on Federated Image Datasets

10.1. Experiment Settings

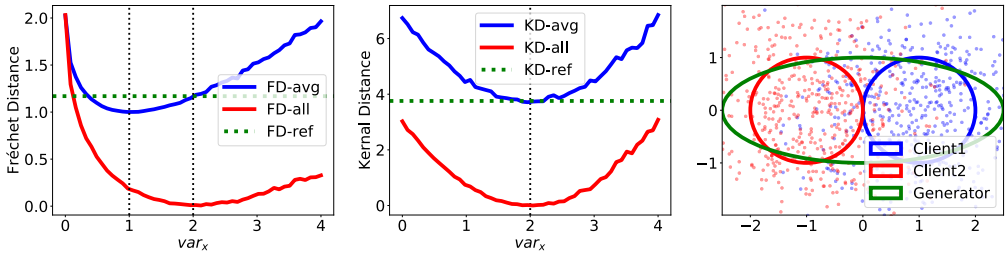
We evaluated our theoretical results on standard image datasets. In our experiments, we simulated heterogeneous federated learning experiments consisting of non-i.i.d. data at different clients: for CIFAR-10 [17], we considered 10 clients, each owning samples exclusively from a single class of the image dataset. Therefore, every client's dataset contains images having the same label. Similar to the federated



(a) Client 1 w/ glasses (b) Client 2 w/ glasses (c) Client 3 w/ glasses

(d) Client 4 w/ head accessories (e) Client 5 w/ head accessories (f) Client 6 w/ head accessories

Figure 14. Illustration of random samples from clients utilized in distributed optimization.



(a) (b) (c)

Figure 15. Experimental results of Gaussian mixture dataset. (a): The optimal var_x parameters are different under FD-avg and FD-all evaluations. (b): Distance between KD-avg and KD-all remains the same.(c): the clients' and generator's samples.

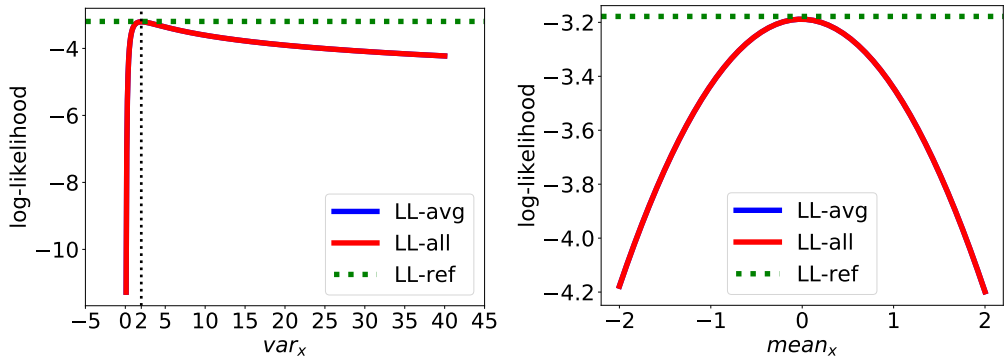
CIFAR10, federated CIFAR100 and federated ImageNet-32 are conducted by grouping samples from a each class.

Neural Net-based Generators. We have trained WGAN-GP [29] and DDPM [12] in a federated learning setting by utilizing FedAvg approach [20]. The experiment protocols for WGAN-GP and DDPM are copied from original works. The communication interval of FedAvg is set as 160 iterations for both WGAN-GP and DDPM. We have tried different communication intervals for both models. The communication frequency will affect model performance but have no influence on the conclusions in the main part of our paper.

Perfect Data-simulating Generators. In our CIFAR-10 experiments, we also simulated and evaluated an "ideal generator" capable of perfectly replicating all samples belonging to the 'airplane' class in CIFAR10. In this scenario, the samples "generated" by the ideal generator exhibit impeccable fidelity but lack diversity since no samples from other categories can be produced.

10.2. FD-based and KD-based Evaluation of Generative Models.

We evaluated the generative models according to FD-all, FD-avg, KD-all, and KD-avg as defined in Section 4. In several cases, we observed that FD-all / FD-avg could assign inconsistent rankings to the generators. Specifically, we computed FD-all and FD-avg for the ideal 'airplane'-class-based generator and neural net-based DDPM generators under the distributed CIFAR10 setting. We present some examples generated from the two generators in Figure 17 and report their scores according to the four metrics. The results suggest that FD-avg assigns a considerably lower score to the ideal 'airplane'-based generator, whose images preserve perfect details but lack diversity in image categories. Conversely, FD-all assigns a relatively lower value to the DDPM model because its images pos-



(a) G_{var_x} (b) G_{mean_x}

Figure 16. Evaluation of synthetic Gaussian data with the aggregate log-likelihood scores.

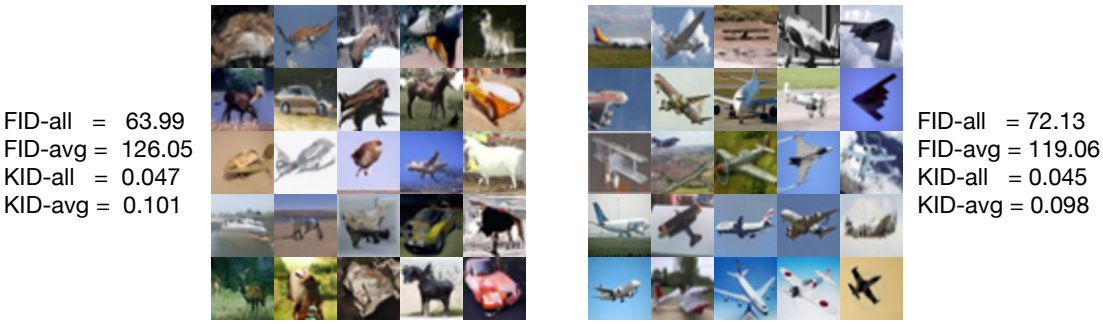


Figure 17. Left: Images generated by a generative model obtain a lower FD-all. Right: Images from real datasets with class 'airplane' obtain a lower FD-avg. FD-avg and FD-all lead to inconsistent rankings, while KD-avg and KD-all result in the same ranking.

sess greater diversity. On the other hand, we also observed that KD-avg and KD-all give consistent rankings. Both of them led to the evaluation that the ideal plane generator is slightly better than the DDPM generator. In our implementation of KD-scores, we utilized the standard implementation of KD measurement from data with a polynomial kernel, $k(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{\mathbf{d}}\mathbf{x}^\mathbf{T}\mathbf{y} + \mathbf{1}\right)^\mathbf{3}$, where d is the dimension of feature vector. We note that our theoretical finding on the evaluation consistency under KD-all and KD-avg applies to every kernel similarity function. We also test KD-scores with a Gaussian RBF kernel $k_\sigma^{\text{rbf}}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\mathbf{1}}{\mathbf{2}\sigma^\mathbf{2}}\|\mathbf{x} - \mathbf{y}\|^\mathbf{2}\right)$ as formulated in [2], where we chose $\sigma = \sqrt{d}$ in the experiments. For images generated by diffusion model KD^{rbf}-all gives $4.277e^{-3}$ while KD^{rbf}-avg gives $4.295e^{-3}$. And for the airplane images in CIFAR10, KD^{rbf}-all gives $4.283e^{-3}$ while KD^{rbf}-avg gives $4.301e^{-3}$. The results indicate that for Gaussian RBF kernel k_σ^{rbf}, KD^{rbf}-all and KD^{rbf}-avg still gives consistent results. In this case, the KD^{rbf}-based evaluation suggests the images sampled from the diffusion model have higher quality than the set of airplane images in the CIFAR10 dataset.

10.3. Evaluate Sequence of Net-based Generator on Federated CIFAR10

We trained the WGAN-GP[9] generative models multiple times using different random states, and we set different training lengths for every training procedure. We saved the models at different checkpoints every 10 epochs, which is common in training generative models to select the best-performing saved model according to an evaluation metric.

Our numerical results suggest that the gap between KD-all and KD-avg remains constant and hence they lead to the same rankings of the generative models. Here, we conducted our evaluations on all the generative models instances as previously described, and the results are visualized in the left sub-figure of Figure 18. These findings reveal that all distinct generators consistently exhibit a uniform gap between KD-avg and KD-all. Consequently, our results indicate that the rankings established by KD-avg consistently align with those of KD-all in distributed learning settings.

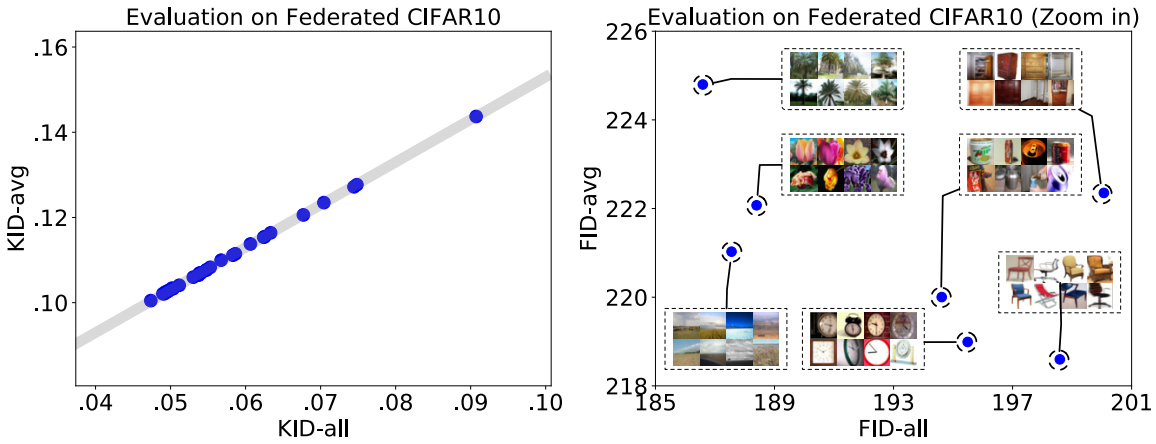


Figure 18. Left: KD evaluations of WGAN-GP checkpoints on federated CIFAR10. Right: FD-based evaluations on federated CIFAR-10.

10.4. Evaluate CIFAR100 Generator on Federated CIFAR10

To further experiment the ranking of generative models according to the discussed aggregate scores, we extracted samples from each class of CIFAR-100 and treated them as the output of one hundred distinct generators, each corresponding to a single class. By assessing these generators on the federated CIFAR-10 dataset, we obtained one hundred pairs of FD-avg / FD-all values, and a subset of these pairs with inconsistent rankings according to FD-all/FD-avg is visualized in the right of Figure 18. The complete set of evaluation results is available on Table 3. These results further highlight that the rankings provided by FD-all and FD-avg can exhibit inconsistencies in the context of distributed learning. Such inconsistencies could pose a challenge when selecting from a series of checkpoints or model architectures during the training of generative models in distributed learning scenarios, where a distributed computation of FD-all is more challenging than obtaining FD-avg due to privacy considerations.

10.5. Evaluate CIFAR100 on Federated ImageNet-32

We expand the evaluation of CIFAR100 to Federated ImageNet-32 dataset. Similarly, we extracted samples from each class of CIFAR-100 and treated them as the output of one hundred distinct generators, each corresponding to a single class. We also keep the first one hundred classes of ImageNet-32 and simulate one hundred clients. Each client hold all images (∼1300) from a single class. We evaluate all the generators on Federated ImageNet-32 and the result is shown in Figure 19. The ranks provided by FD-avg and FD-all is inconsistent in a much more complex distributed learning setting.

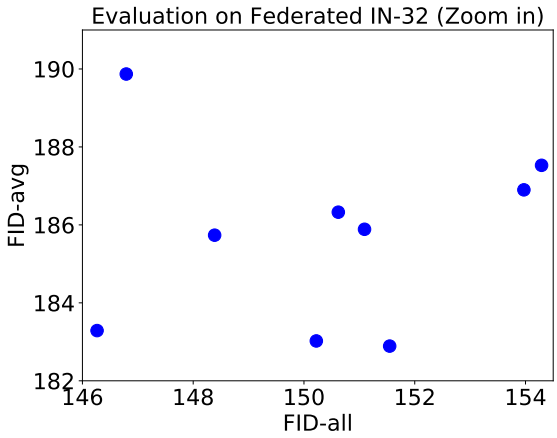


Figure 19. Evaluate CIFAR100 on Federated ImageNet-32.

11. Evaluation on Variance-Limited Federated Datasets

11.1. Experiment Setting

In the federated learning literature, it is relatively common that each client possesses only a small portion of the collective dataset, and the data diversity within each client's holdings is significantly constrained. To illustrate, consider the case of smartphone users who exclusively own pictures of themselves, all of which share remarkable similarity. Nevertheless, in a network comprising millions of users, the overall dataset's distribution still exhibits significant variance. In such scenarios, our theoretical framework suggests that the disparity between FD-all and FD-avg can become more pronounced. To experiment the effect of such distribution heterogeneity, we simulated and evaluated generative models under variance-limited federated datasets. To obtain a variance-limited federated dataset, for each class

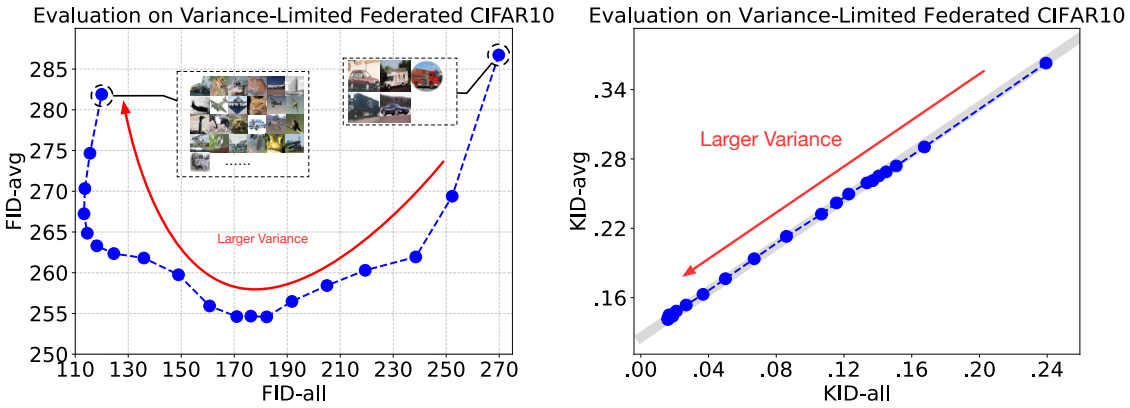


Figure 20. Evaluation on Variance-Limited Federated CIFAR10.

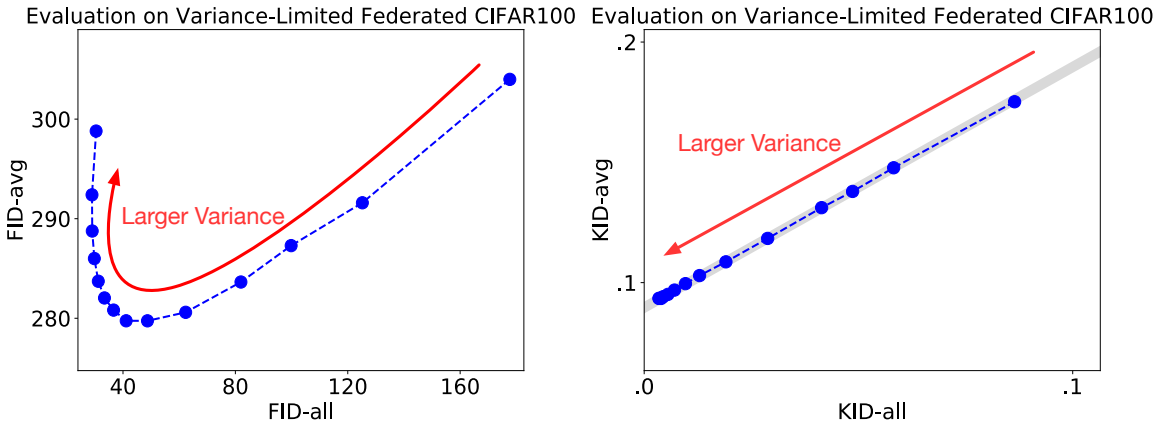


Figure 21. Evaluation on Variance-Limited Federated CIFAR100.

in the image dataset, we kept only a single image and its K-nearest neighbors. To find the K nearest neighbors, we used the L_2-distance in the Inception-V3 2048-dimensional semantic space. It is worth noting that our experimentation has shown that varying K within the range of 5 to 100 does not alter the core conclusions. This approach effectively mimics scenarios where each client's data has limited variance. We simulated the variance-limited federated learning setting for CIFAR-10, CIFAR-100 and a 32×32 version of ImageNet (IN-32). For CIFAR-10 and CIFAR-100, we utilized all the classes in the dataset and for IN-32 we utilized the first 100 classes. We chose $K = 20$ in the experiments. Intuitively, a larger K leads to a more significant intra-client variance.

Variance-controlled Generators. To simulate a generator, we initiate the process by randomly selecting a sample from the dataset. We then gather its M-nearest neighbors from the original dataset (w/o federated learning setting). We consider this subset of samples as a set of generated samples generated by a generator denoted by G_M. By increasing the value of M, we generated a sequence of generators with progressively higher variance values. We tried the M range from 100 to 50000. We evaluated all the generative models, denoted as G_M with the chosen M values, using the Variance-Limited Federated datasets.

11.2. Results on Variance-Limited Federated CIFAR10

The evaluation results on CIFAR10 are shown in Figure 20. Our findings reveal a distinct pattern in the behavior of FD-avg and FD-all as generator variance varies while the distance between KD-avg and KD-all remains the same. Our numerical results highlight the impact of the choice of FD-all and FD-avg on model rankings in federated learning settings with limited intra-client variance. , which can be broadly categorized into three phases.

11.3. Results on Variance-Limited Federated CIFAR100

Similar to the experiments on CIFAR10, we have also applied the variance-limited federated dataset setting to CI-

FAR100. We keeps K=20 images in each class. For variance-controlled generators, we select a sample from original CIFAR100 and gather the M-nearset neighbors. The range of M keeps the same with that in the previous subsection. We show the results in Figure 21. The results still support our main claims: FD-avg and FD-all gives inconsistent results while KD-avg and KD-all give the same.

11.4. Results on Variance-Limited Federated IN32

Results on ImageNet-32 are illustrated in Figure 22. We further plot the relationship between ranking given by FD-scores and KD-scores in Figure 23.

11.5. The Effect of Intra-Client Variance

In the main body of this paper, we choose $K = 20$ when we conduct the variance-limited federated CIFAR10 dataset. Hyper-parameter K controls the intra-client variance, the larger the K the larger the variance. The number of K will not affect the key conclusion. We prove this claim by conducting an ablation study on hyper-parameter K. The K is selected from $\{5,10,20,50\}$ in our experiment. The results are illustrated in Figure 24. Each of these figure gives a U-shape curve, which indicates that the rankings given by FD-all and FD-avg are highly inconsistent, especially when the intra-client variance and inter-client variance are mismatched.

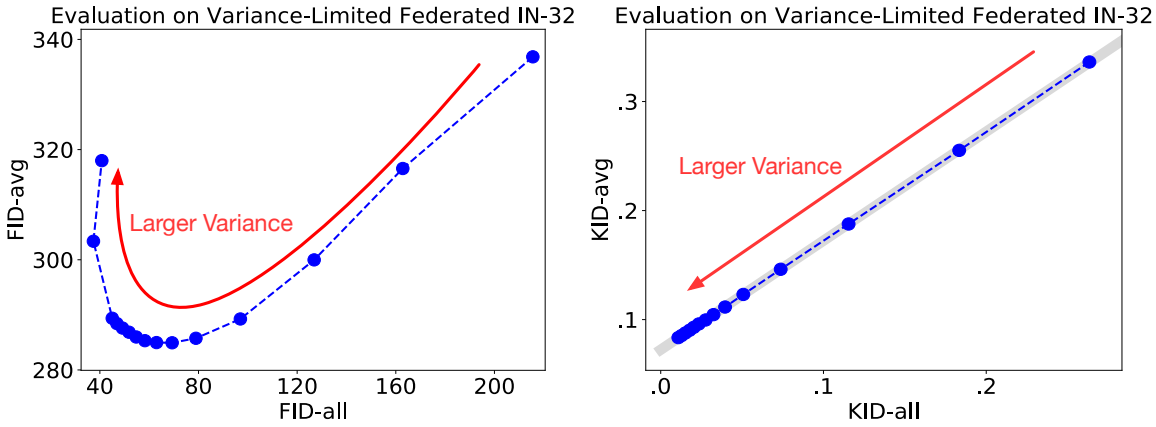


Figure 22. FD and KD-based Evaluations of variance-controlled generators on variance-limited federated ImageNet-32.

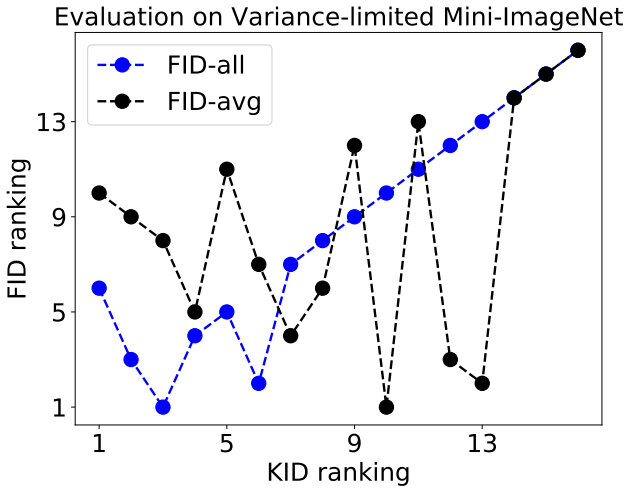


Figure 23. Comparing FD-based and KD-based rankings of variance-limited federated Mini-ImageNet-based simulated generative models. The lower the rank is, the better.

Class	FD-all	FD-avg	KD-all	KD-avg	Class	FD-all	FD-avg	KD-all	KD-avg
0	267.4	285.9	0.201	0.253	25	142.4	173.2	0.067	0.116
1	173.2	205.5	0.109	0.165	26	139.4	175.5	0.063	0.114
2	151.8	185.4	0.072	0.123	27	114.5	153.7	0.054	0.102
3	124.3	162.2	0.047	0.100	28	182.0	205.0	0.112	0.160
4	117.0	156.1	0.049	0.100	29	143.3	185.2	0.083	0.133
5	157.6	185.2	0.088	0.138	30	147.4	179.8	0.084	0.134
6	142.1	179.7	0.062	0.115	31	160.4	193.6	0.108	0.158
7	144.9	179.2	0.066	0.113	32	115.4	156.4	0.032	0.085
8	155.0	187.7	0.085	0.135	33	146.9	181.2	0.077	0.126
9	180.7	203.6	0.106	0.156	34	127.5	163.4	0.068	0.117
10	183.2	208.8	0.099	0.151	35	150.4	184.7	0.080	0.131
11	151.9	184.9	0.073	0.127	36	142.5	176.5	0.070	0.126
12	126.1	163.1	0.056	0.108	37	126.7	162.9	0.066	0.118
13	127.0	159.4	0.058	0.112	38	110.5	151.3	0.050	0.100
14	145.8	182.9	0.071	0.126	39	233.5	257.2	0.145	0.196
15	124.2	165.2	0.054	0.107	40	158.9	187.6	0.078	0.128
16	190.3	213.6	0.112	0.161	41	151.1	181.0	0.072	0.126
17	161.5	192.8	0.117	0.165	42	137.2	174.7	0.071	0.122
18	142.7	179.8	0.069	0.118	43	152.0	186.0	0.090	0.139
19	112.4	154.1	0.047	0.101	44	126.7	165.8	0.050	0.103
20	194.2	215.9	0.123	0.175	45	135.4	173.0	0.059	0.111
21	164.8	194.9	0.099	0.148	46	155.1	187.1	0.080	0.133
22	205.1	227.0	0.126	0.180	47	174.2	203.4	0.118	0.169
23	191.1	219.7	0.130	0.186	48	145.0	175.3	0.077	0.129
24	174.9	202.8	0.105	0.155	49	164.8	194.6	0.110	0.164
50	113.8	154.0	0.045	0.095	75	151.7	184.3	0.096	0.145
51	150.4	184.6	0.073	0.125	76	151.5	183.5	0.068	0.118
52	195.3	222.1	0.167	0.218	77	139.0	174.9	0.066	0.117
53	279.7	299.1	0.217	0.270	78	197.3	228.3	0.130	0.182
54	170.2	201.3	0.098	0.149	79	138.9	174.8	0.062	0.114
55	104.7	146.3	0.034	0.086	80	111.1	149.9	0.043	0.095
56	144.2	178.5	0.075	0.126	81	131.7	165.3	0.068	0.119
57	192.5	219.9	0.115	0.163	82	197.6	227.2	0.123	0.178
58	131.5	161.0	0.067	0.121	83	202.6	230.6	0.122	0.173
59	149.7	183.7	0.093	0.144	84	144.5	177.3	0.065	0.111
60	188.0	216.0	0.144	0.197	85	123.0	160.7	0.079	0.125
61	249.2	270.6	0.179	0.229	86	168.0	193.0	0.083	0.135
62	202.1	230.6	0.133	0.184	87	170.5	196.0	0.098	0.152
63	140.6	175.3	0.071	0.120	88	133.4	170.6	0.068	0.120
64	118.7	156.1	0.049	0.100	89	122.3	158.1	0.059	0.112
65	102.2	142.2	0.024	0.077	90	110.7	148.3	0.041	0.093
66	121.7	159.1	0.054	0.105	91	124.0	160.7	0.048	0.098
67	132.5	167.8	0.063	0.115	92	175.4	206.2	0.096	0.149
68	139.7	173.1	0.073	0.123	93	129.7	166.3	0.054	0.109
69	143.2	176.0	0.068	0.121	94	213.4	235.2	0.162	0.212
70	178.4	209.5	0.095	0.148	95	154.7	185.2	0.084	0.133
71	169.4	199.0	0.120	0.167	96	147.8	181.7	0.090	0.138
72	114.1	155.6	0.043	0.094	97	137.1	171.4	0.068	0.119
73	137.9	170.8	0.071	0.124	98	157.1	188.6	0.082	0.134
74	124.7	162.1	0.061	0.108	99	204.9	233.9	0.143	0.192

Table 3. Full evaluation of CIFAR100 on Federated CIFAR10.

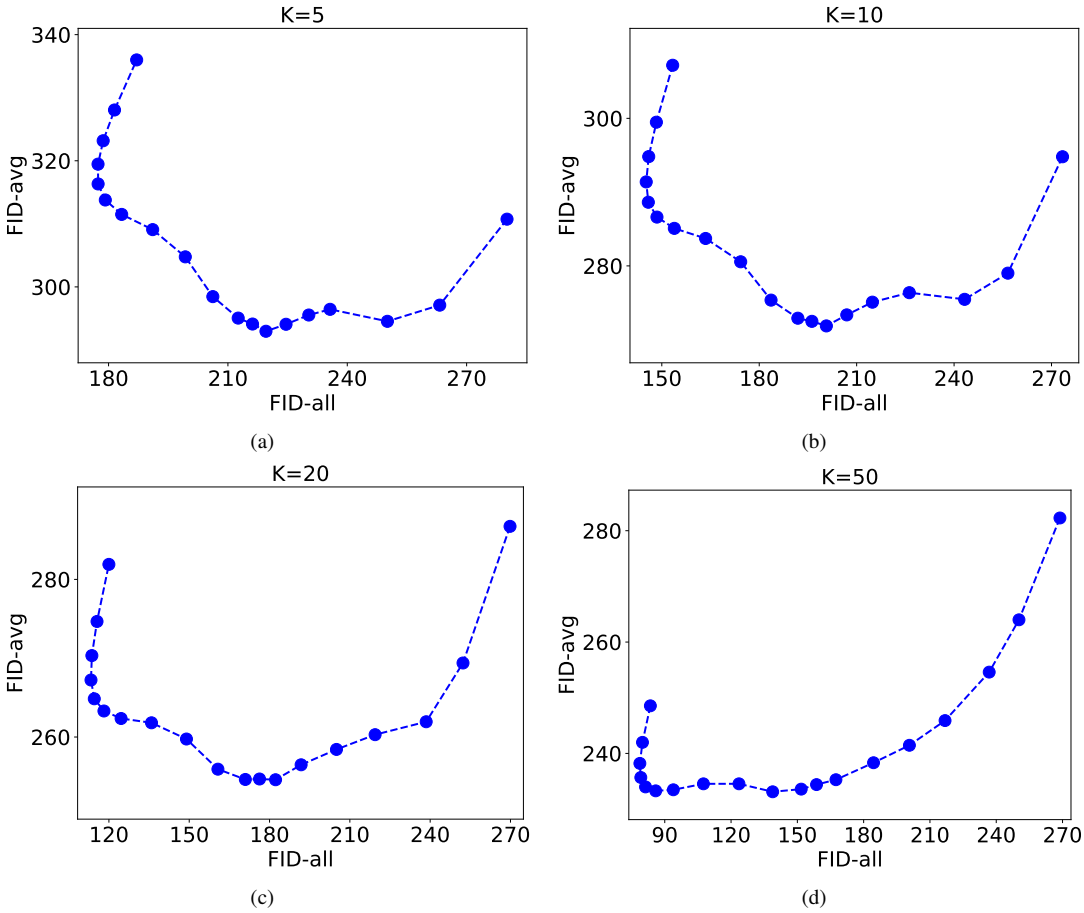


Figure 24. Ablation study on hyper-paramter K.

12. Proofs

12.1. Proof of Theorem 1

To show this theorem, we note that if $\phi(X)$ is the kernel feature map for kernel k used to define the KD distance, i.e. $k(x,y) = \langle \phi(x), \phi(y) \rangle$ is the inner product of the feature maps applied to x, y, then it can be seen that the kernel-k-based MMD distance can be written as

$$\begin{aligned}\mathrm{MMD}\big(P_X, P_G\big) :=& \mathbb{E}_{X,X' \sim P_X}\big[k(X,X')\big] + \mathbb{E}_{Y,Y' \sim P_G}\big[k(Y,Y')\big] - 2\,\mathbb{E}_{X \sim P_X, Y \sim P_G}\big[k(X,Y)\big] \\ =& \Big\| \mathbb{E}\big[\phi(X)\big] - \mathbb{E}\big[\phi(Y)\big] \Big\|^2 .\end{aligned}$$

Therefore, following the definition of KD-avg, we can write

$$\begin{aligned}\mathrm{KD}_{\mathrm{avg}}\Big(P_{X_1}, \ldots, P_{X_k}\,;\, P_G\Big) \;&:=\; \sum_{i=1}^{k} \lambda_i \mathrm{KD}\big(P_{X_i}, P_G\big) \\ &= \sum_{i=1}^{k} \lambda_i \mathrm{MMD}_\phi\big(P_{X_i}, P_G\big) \\ &\overset{(a)}{=} \sum_{i=1}^{k} \lambda_i \Big\| \mathbb{E}\big[\phi(X_i)\big] - \mathbb{E}\big[\phi(G(Z))\big] \Big\|^2 \\ &\overset{(b)}{=} \big\| \mathbb{E}\big[\phi(\widehat{X})\big] - \mathbb{E}\big[\phi(G(Z))\big] \big\|^2 + \sum_{i=1}^{k} \Big[\lambda_i \big\| \mathbb{E}\big[\phi(X_i)\big] - \mathbb{E}\big[\phi(\widehat{X})\big] \big\|^2 \Big] \\ &\overset{(c)}{=} \mathrm{MMD}_\phi\big(\widehat{P}_X, P_G\big) + \sum_{i=1}^{k} \Big[\lambda_i \mathrm{MMD}_\phi\big(\widehat{P}_X, P_{X_i}\big) \Big] \\ &\overset{(d)}{=} \mathrm{KD}\big(\widehat{P}_X, P_G\big) + \sum_{i=1}^{k} \Big[\lambda_i \mathrm{KD}\big(\widehat{P}_X, P_{X_i}\big) \Big] \\ &\overset{(e)}{=} \mathrm{KD}_{\mathrm{all}}\Big(P_{X_1}, \ldots, P_{X_k}\,;\, P_G\Big) + \sum_{i=1}^{k} \lambda_i \mathrm{KD}\big(\widehat{P}_X, P_{X_i}\big).\end{aligned}$$

In the above, (a) and (c) follow from the feature-map-based formulation of the MMD distance. (b) is the consequence of the fact that $\|\cdot\|$ is the norm in a reproducing kernel Hilbert space and for $\widehat{X}$ distributed as $\widehat{P}_X = \sum_{i=1}^{k} \lambda P_{X_i}$ we know that $\mathbb{E}\big[\phi(\widehat{X})\big]$ is the weighted barycenter of the individual mean vectors $\mathbb{E}\big[\phi(X_1)\big], \ldots, \mathbb{E}\big[\phi(X_k)\big]$. (d) is based on the definition of KD. Finally, (e) follows from the definition of KD-all, which completes the proof.

12.2. Proof of Theorem 2

1. Note that according to the definition,

$$\mathrm{FD}_{\mathrm{all}}\Big(P_{X_1}, \ldots, P_{X_k}\,;\, P_G\Big) \;=\; \mathrm{FD}\big(\sum_{i=1}^{k} \lambda_i P_{X_i}, P_G\big).$$

Since the FD score depends only on the the mean and covariance parameters in the Embedding-based semantic space, we can replace $\sum_{i=1}^{k} \lambda_i P_{X_i}$ with any other distribution that shares the same mean and covariance parameters, and the FD value will not change. Observe that given mean parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, the Embedding-based mean of $\sum_{i=1}^{k} \lambda_i P_{X_i}$ will

be $\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{k} \lambda_i \boldsymbol{\mu}_i$. Therefore, the Embedding-based covariance matrix of $\sum_{i=1}^{k} \lambda_i P_{X_i}$ follows from

$$\begin{aligned}\sum_{i=1}^{k} \lambda_i \mathbb{E}_{P_i}\left[(X_i - \widehat{\boldsymbol{\mu}})(X_i - \widehat{\boldsymbol{\mu}})^\top\right] &= \sum_{i=1}^{k} \lambda_i \Big[C_i + (\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})^\top\Big] \\ &= \sum_{i=1}^{k} \lambda_i \Big[C_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top\Big] - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top \\ &= \widehat{C}.\end{aligned}$$

Therefore, since we assume $\widehat{X}$ has the Embedding-based mean and covariance $\widehat{\boldsymbol{\mu}}$ and $\widehat{C}$, the proof of this part is complete.

2. According to the definition, FD-avg can be written as

$$\mathrm{FD}_{\mathrm{avg}}\Big(P_{X_1}, \ldots, P_{X_k}\,;\, P_G\Big) \;:=\; \sum_{i=1}^{k} \lambda_i \mathrm{FD}\big(P_{X_i}, P_G\big).$$

Therefore, we have

$$\begin{aligned}&\mathrm{FD}_{\mathrm{avg}}\Big(P_{X_1}, \ldots, P_{X_k}\,;\, P_G\Big) \\ &\overset{(a)}{=} \sum_{i=1}^{k} \lambda_i W_2^2\Big(\mathcal{N}(\boldsymbol{\mu}_i, C_i), \mathcal{N}(\boldsymbol{\mu}_G, C_G)\Big) \\ &\overset{(b)}{=} \sum_{i=1}^{k} \lambda_i \Big[\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_G\|_2^2 + \mathrm{Tr}\big(C_i + C_G - (C_i C_G)^{1/2}\big)\Big] \\ &= \sum_{i=1}^{k} \Big[\lambda_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_G\|_2^2\Big] + \sum_{i=1}^{k} \Big[\lambda_i \mathrm{Tr}\big(C_i + C_G - (C_i C_G)^{1/2}\big)\Big] \\ &\overset{(c)}{=} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_G\|_2^2 + \sum_{i=1}^{k} \Big[\lambda_i \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i\|_2^2\Big] \\ &\qquad + \mathrm{Tr}\big(C_G + \widehat{C} - (C_G \widehat{C})^{1/2}\big) + \sum_{i=1}^{k} \Big[\lambda_i \mathrm{Tr}\big(C_i + \widehat{C} - (C_i \widehat{C})^{1/2}\big)\Big] \\ &= \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_G\|_2^2 + \mathrm{Tr}\big(C_G + \widehat{C} - (C_G \widehat{C})^{1/2}\big) \\ &\qquad + \sum_{i=1}^{k} \Big[\lambda_i \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i\|_2^2 + \lambda_i \mathrm{Tr}\big(C_i + \widehat{C} - (C_i \widehat{C})^{1/2}\big)\Big] \\ &= \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_G\|_2^2 + \mathrm{Tr}\big(C_G + \widehat{C} - (C_G \widehat{C})^{1/2}\big) \\ &\qquad + \sum_{i=1}^{k} \lambda_i \Big[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i\|_2^2 + \mathrm{Tr}\big(C_i + \widehat{C} - (C_i \widehat{C})^{1/2}\big)\Big] \\ &\overset{(d)}{=} \mathrm{FD}(P_{\widehat{X}}, P_G) + \sum_{i=1}^{k} \lambda_i \mathrm{FD}(P_{\widehat{X}}, P_{X_i}).\end{aligned}$$

In the above, (a) follows from the Wasserstein-based definition of FD distance. (b) comes from the well-known closed-form expression of the 2-Wasserstein distance between Gaussian distributions [36]. (c) is the result of applying the weighted barycenter of vector $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ that can be seen to be $\widehat{\boldsymbol{\mu}}$ and the weighted barycenter of positive semi-definite covariance matrices $C_1, \ldots, C_k$ that has been shown to be the unique matrix $\widehat{C}$ that solves the equation $\widetilde{C} = \sum_{i=1}^{k} \lambda_i \big(\widetilde{C}^{1/2} C_i \widetilde{C}^{1/2}\big)^{1/2}$ [24, 27]. (d) is the direct consequence of the Wasserstein-based definition of the FD distance and the closed-form expression of the 2-Wasserstein distance between Gaussians. Therefore, the proof is complete.

12.3. Proof of Proposition 1

Consider the FD-all-minimizing parameters in Theorem 2 resulting in

$$\begin{aligned}\mathrm{FD}_{\mathrm{all}}\Big(P_{X_1},\ldots,P_{X_k}\,;\,P_{\widehat{G}}\Big) &= \mathrm{FD}\big(\mathcal{N}(\widehat{\boldsymbol{\mu}},\widehat{C}),\mathcal{N}(\boldsymbol{\mu}_{\widehat{G}},C_{\widehat{G}})\big)\\ &= \big\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\widehat{G}}\big\|_2^2+\mathrm{Tr}\Big(\widehat{C}+C_{\widehat{G}}-2\big(\widehat{C}C_{\widehat{G}}\big)^{1/2}\Big).\end{aligned}$$

Note that since we assume the number of clients k is less than the dimension of the embedding, there exists a unit-norm vector $\boldsymbol{\beta}$ ($\|\boldsymbol{\beta}\|_2=1$) in the embedding space that is orthogonal to all mean vectors $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k$ and hence to their mean $\widehat{\boldsymbol{\mu}}=\frac{1}{k}\sum_{i=1}^k\boldsymbol{\mu}_i$. Given $u=\mathrm{Tr}\big(\sum_{i=1}^k\lambda_i\big(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top-\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top\big)\big)$, we then consider the generator G' with the following mean and covariance parameters:

$$\boldsymbol{\mu}_{G'}=\widehat{\boldsymbol{\mu}}+\sqrt{u}\boldsymbol{\beta},\qquad C_{G'}=C_{\widehat{G}}-\sum_{i=1}^k\lambda_i\big(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top-\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top\big)=\sum_{i=1}\lambda_iC_i.$$

We claim that the generators $\widehat{G}$ and G' lead to the same client-based FD scores as for every i

$$\begin{aligned}\mathrm{FD}\big(P_{X_i},P_{G'}\big) &= \mathrm{FD}\big(\mathcal{N}(\boldsymbol{\mu}_i,C_i),\mathcal{N}(\boldsymbol{\mu}_{G'},C_{G'})\big)\\ &= \big\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_{G'}\big\|_2^2+\mathrm{Tr}\Big(C_i+C_{G'}-2\big(C_iC_{G'}\big)^{1/2}\Big)\\ &= \big\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_{\widehat{G}}\big\|_2^2+u+\mathrm{Tr}\Big(C_i+C_{G'}-2\big(C_iC_{\widehat{G}}\big)^{1/2}\Big)-u\\ &= \big\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_{\widehat{G}}\big\|_2^2+\mathrm{Tr}\Big(C_i+C_{G'}-2\big(C_iC_{\widehat{G}}\big)^{1/2}\Big)\\ &= \mathrm{FD}\big(P_{X_i},P_{\widehat{G}}\big).\end{aligned}$$

On the other hand, for the FD-all of G' we have

$$\begin{aligned}\mathrm{FD}_{\mathrm{all}}\Big(P_{X_1},\ldots,P_{X_k}\,;\,P_{G'}\Big) &= \mathrm{FD}\big(\mathcal{N}(\widehat{\boldsymbol{\mu}},\widehat{C}),\mathcal{N}(\boldsymbol{\mu}_{G'},C_{G'})\big)\\ &= \big\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{G'}\big\|_2^2+\mathrm{Tr}\Big(\widehat{C}+C_{G'}-2\big(\widehat{C}C_{G'}\big)^{1/2}\Big)\\ &= \big\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\widehat{G}}\big\|_2^2+u\mathrm{Tr}\Big(\widehat{C}+C_{\widehat{G}}-2\big(\widehat{C}C_{\widehat{G}}\big)^{1/2}\Big)+u\\ &= \big\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\widehat{G}}\big\|_2^2+\mathrm{Tr}\Big(\widehat{C}+C_{\widehat{G}}-2\big(\widehat{C}C_{\widehat{G}}\big)^{1/2}\Big)\\ &\quad+2\mathrm{Tr}\big(\sum_{i=1}^k\lambda_i\big(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top-\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top\big)\big)\\ &= \mathrm{FD}_{\mathrm{all}}\Big(P_{X_1},\ldots,P_{X_k}\,;\,P_{\widehat{G}}\Big)\\ &\quad+2\mathrm{Tr}\big(\sum_{i=1}^k\lambda_i\big(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top-\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top\big)\big)\end{aligned}$$

Therefore, Proposition 1's proof is complete.