A. Differential Privacy Analysis

We analyze the differential privacy of our proposed methods, adopting the same definition as [34] for differential privacy in randomized mechanisms. We show that our proposed method satisfies $(0, \delta)$ differential privacy or $(0, \delta)$-DP for short.

Definition A.1 ((ϵ, δ)-DP)**.** *A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP if for all measurable sets $\mathcal{S} \subset \mathcal{R}$ and for any two adjacent databases $\mathcal{C}$ and $\mathcal{C}' \in \mathcal{X}$,*

$$P(\mathcal{M}(\mathcal{C}) \in \mathcal{S}) \leq e^{\epsilon} P(\mathcal{M}(\mathcal{C}') \in \mathcal{S}) + \delta$$

Since we focus on the client level perspective, the databases $\mathcal{C}$ and $\mathcal{C}'$ here are the sets of clients, which differ on one client only, c and c', *i.e.*,

$$\begin{aligned} \mathcal{C} &= c \cup \mathcal{C}_0, \\ \mathcal{C}' &= c' \cup \mathcal{C}_0. \end{aligned} \tag{4}$$

Here, we denote the distributions of the datasets D and D' of the two client sets $\mathcal{C}$ and $\mathcal{C}'$ as $P_D(X)$ and $P_{D'}(X)$. Assume both clients start training their models, on their local datasets, starting from the same initial parameter W, e.g. the global model. If their datasets having different distributions, both clients will obtain two different models after local training, which have different parameter distributions. We denote the two parameter distributions as $P_{\mathcal{C}}(W)$ and $P_{\mathcal{C}'}(W)$. For simplicity, we assume the model training is a stochastic process estimating the following posterior distribution according to the Bayes' rule,

$$P(W|X) \propto P(X|W)P_0(W),$$

where $P_0(W)$ is the prior distribution of W. Since each client trains on the same model architecture, the likelihood model $P(W|X)$ will be the same for all clients. It is also reasonable to use the same prior distribution for every client.

Assumption A.1. *The total variation distance (TV) between the distributions of any two different augmented client datasets are less than δ: $TV(P_D(X), P_{D'}(X)) \leq \delta$.*

To verify the assumption A.1, we denote the distribution of generated data as G, and the i-th client's dataset is the union of the generated data and the raw data, and the distribution of this combined dataset is denoted as P_i. According to the definition of TV distance and its triangle inequality, given an arbitrary δ, we can always generate large enough samples such that $TV(G, P_i)$ is smaller than $\delta/2$. Thus for any two clients, we have $TV(P_j, P_i) \leq TV(P_j, G) + TV(P_i, G) \leq \delta/2 + \delta/2 = \delta$. As a result, the assumption A.1 is reasonable. With the above assumption, we use the data processing inequality stated in Lemma A.1 to derive the TV distance between $P_{\mathcal{C}}(W)$ and $P_{\mathcal{C}'}(W)$.

Lemma A.1. *(Theorem 6.2 in [2]) Consider a channel that produces Y given X based on the law $P_{Y|X}$ (illustrated in Figure 6). If P_Y is the distribution of Y when X is generated by P_X and Q_Y is the distribution of Y when X is generated by Q_X, then for any f-divergence $D_f(\cdot\|\cdot)$,*

$$D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X)$$

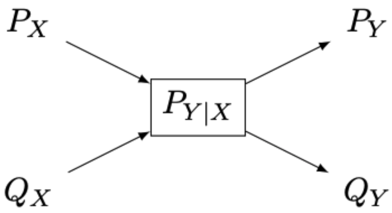

Figure 6: Data processing inequality

Theorem A.2. *Federated learning with zero-shot data augmentation satisfies the differential privacy $(0, \delta)$-DP.*

Proof. Since the total variation distance is an instance of f-divergence [2], applying Lemma A.1, we obtain

$$TV(P_{\mathcal{C}}(W), P_{\mathcal{C}'}(W)) \leq TV(P_D(X), P_{D'}(X)) \leq \delta.$$

In federated learning, we perform model aggregation, denoted as W_{agg}, as

$$W_{agg} = \frac{1}{n}W + \frac{n-1}{n}W_0$$

where W_0 is the parameter aggregated on the set of other clients $\mathcal{C}_0$ (as defined in Eq. 4) and n is the number of clients in $\mathcal{C}$. We denote the two different distributions of W_{agg} in the two models as $P_{\mathcal{C}}(W_{agg})$ and $P_{\mathcal{C}'}(W_{agg})$. Similarly, we can also use the Lemma A.1 to derive that,

$$TV(P_{\mathcal{C}}(W_{agg}), P_{\mathcal{C}'}(W_{agg})) \leq TV(P_{\mathcal{C}}(W), P_{\mathcal{C}'}(W)) \leq \delta$$

Based on the definition of total variation distance, we have

$$\sup_{S \subset R} |P_{\mathcal{C}}(W_{agg} \in S) - P_{\mathcal{C}'}(W_{agg} \in S)| \leq \delta$$

Define the stochastic mechanism M as the projection from the client set to any model parameter $W_{agg} \in \mathcal{R}$. Then the distribution of $M(\mathcal{C})$ and $M(\mathcal{C}')$ are the distributions of W_{agg} and W'_{agg}, respectively. Hence, for any $S \subset R$:

$$P(M(\mathcal{C}) \in S) \leq P(M(\mathcal{C}') \in S) + \delta\ ,$$

which finishes the proof that Fed-ZDA satisfies $(0, \delta)$-DP. □